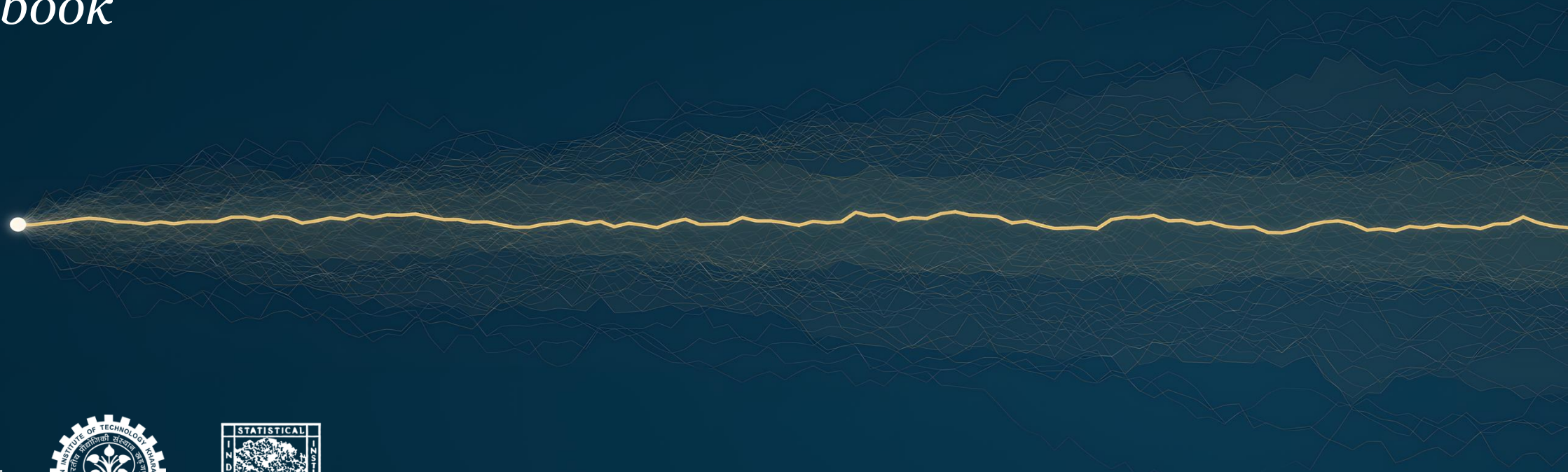
The PGDBA

Data Science

Casebook



IIM C IIT KGP ISI K

Copyright

PGDBA Data Science Casebook, AINA Magazine, Indian Institute of Management Calcutta, 1st edition

Introducing PGDBA's Data Science Casebook

The PGDBA Publishing Team is proud to present the first edition of the Data Science Casebook.

This casebook has been compiled from real cases featured in data science interviews. The transcripts have been adapted to more holistically portray the array of decisions every data scientist must make in real business scenarios that aren't as clean-cut as classroom examples, and these are supplemented by flowcharts to clarify the thought process employed by the interviewee. Also included are overviews of major industries that invest significantly into data science expertise, as well as a primer that covers key concepts that are frequently tested in data analytics interviews.

While this document is in part intended to serve as a basis for interview preparation, we strongly recommend that readers supplement the concepts contained within with mathematically rigorous foundations and hands-on implementations. For any queries, please feel free to reach out to pgdba.aina@email.iimcal.ac.in

We sincerely hope that this casebook proves to be a valuable resource for all data scientists in the making!

Jaganatha Pandiyan
Editor-in-Chief

Srirang Nabar

Sanket Wathore

Index

Market Entry for Electric Scooters – Transcript

Interviewer: Your client wants to introduce electric scooters in tier-2 cities. How should they go about identifying potential sales and profitability?

Candidate: Could you tell me about the existing competitors and EV penetration in tier-2 cities? Also, how does EV pricing compare to regular scooters?

According to the client, about 10% of scooters in tier-2 cities are electric. Two major players control around 60% of the EV market, with the rest fragmented. Electric scooters are approximately 30% more expensive than petrol scooters.

My approach would be to first estimate market attractiveness by sizing the addressable demand, then model the client's likely market share using data-driven methods, and finally combine this with unit economics to estimate profitability.

That sounds like a solid plan. Let's start with estimating the total demand in tier-2 cities.

Given India's population of ~1.4 billion, I'll isolate the ~35% urban segment, which is ~500 million people. Assuming tier-2 cities account for ~30% of this, we get ~150 million as the base population. From here, I'll apply two filters to arrive at the addressable base:

1. **Age –** scooters are predominantly used by those in the 18–50 age segment, which is ~55% of the population, or 80 million people.
2. **Affordability –** individuals with sufficient earning power or access to financing to afford a scooter, which we'll approximate at ~25% of the 18–50 segment, bringing the base to ~20 million people.

That seems reasonable. How will you proceed from here?

I'll split this base into two demand streams.

1. **Replacement Demand –** let's say 60% of this base already owns a scooter and will replace it at end-of-life. With an average lifetime of about 8 years, the simple approximation is that 1/8 of current owners enter the market each year, about 1.5 million people (20 million * 0.6 * (1/8)).
2. **First-time demand –** the remaining 40% are first-time buyers driven by income growth and urbanisation. Spreading this over the next 5 years to allow gradual income-driven adoption gives us an annual demand of 1.6 million (20 million * 0.4 * (1/5)).

Adding these gives a total annual scooter market in tier-2 cities of ~3.1 million units.

How much of this demand is for electric scooters?

Given the current EV penetration of ~10%, this translates to ~0.3 million electric scooters annually.

However, EV adoption is growing due to subsidies and lower running costs. I would model this increasing to ~20–25% over the next few years, rather than keeping it static, assuming charging infrastructure expands and cost parity improves

Let's just focus on the first year. So can the client expect ~0.3 million sales in the first year?

No, that is the total market demand. To estimate the client's share, I would use a data-driven approach. I'd model the client's market share as a function of key drivers using a fractional logit model (since share is bounded in [0,1]), using data from tier-1 cities where the client already operates to estimate coefficients, with features like relative pricing (vs competitors), dealer density, product specifications, marketing spend, and brand awareness. Using these x values and estimated coefficients, we can estimate the client's share in the tier-2 market, assuming the mapping from drivers to share is stable across tiers.

How will you forecast demand for the next 2 - 3 years after launch?

Time-series models like SARIMAX or Prophet require city-wise historical sales data, which tier-2 cities lack because this is a cold start. I'd therefore use cross-sectional approaches that learn from tier-1 cities and generalise across cities rather than across time. I'd use two cold start approaches:

1. **Hierarchical regression with city random intercepts –** fits on pooled tier-1 city data using covariates (for example, income, fuel price, charging density, dealer count). Tier-2 forecasts use the coefficients trained on tier-1 data, with the random intercept set to its prior mean, since the city is unseen
2. **Gradient boosting on cross-sectional features –** captures nonlinear interactions

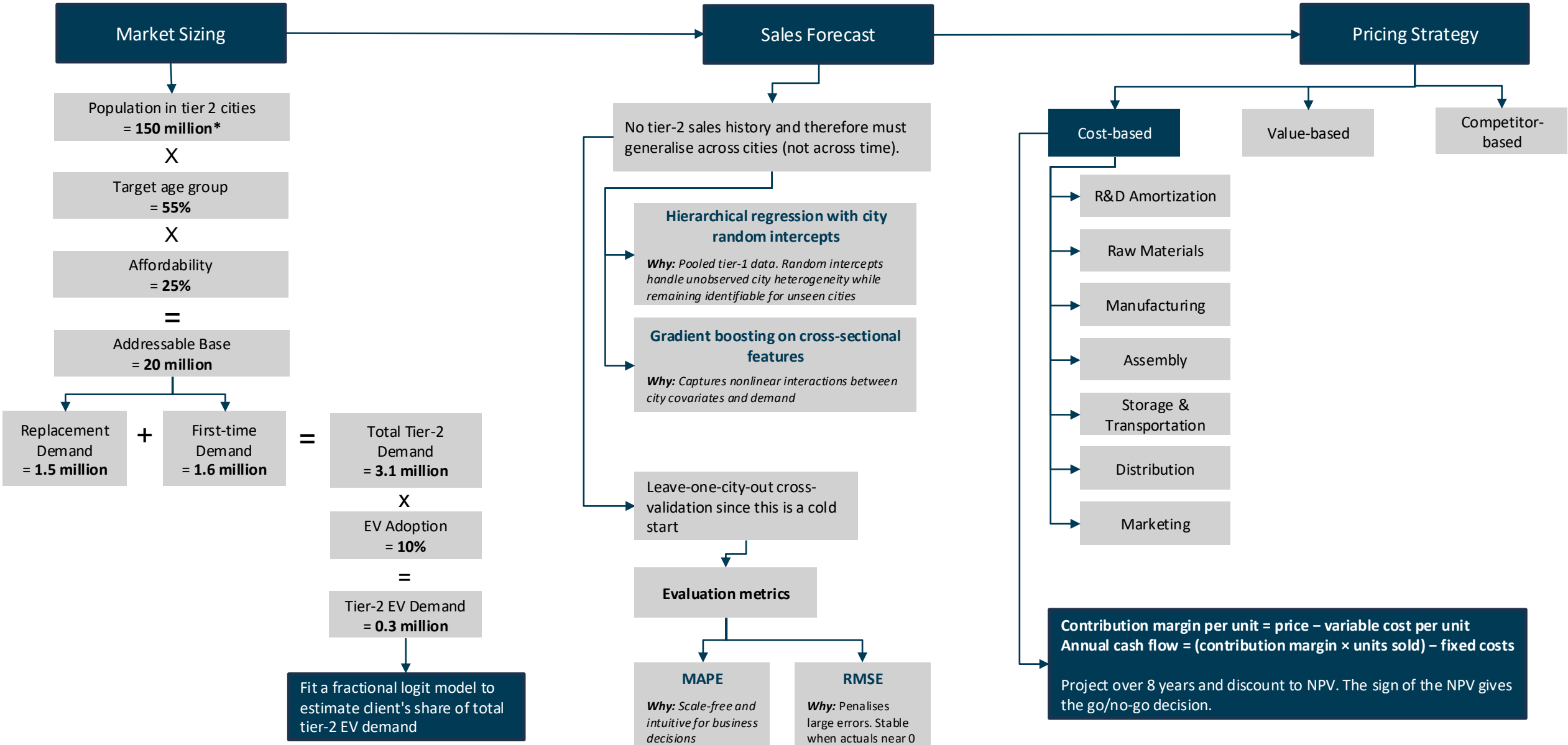
I'd evaluate using MAPE and RMSE with leave-one-city-out cross-validation, since the test is generalising across cities rather than time, and then retrain as tier-2 data accumulates.

Good. How will you approach identifying profitability?

Since this is a new product, I'll break down the cost across the value chain: R&D amortisation, raw material procurement, manufacturing, assembly, storage & transportation, distribution, and marketing. This breakdown gives us the variable cost per unit and annual fixed costs. From there, the contribution margin per unit is price minus variable cost. I'd combine this with the demand forecast to compute a cash flow stream over an 8-year horizon, discounted to NPV. The sign of NPV gives the go/ no-go decision. I would test the decision under bear, base and bull scenarios of adoption and pricing.

Excellent

Market Entry for Electric Scooters – Case Flow



∗ *Population in tier 2 cities = population of India ∗ urban % ∗ tier 2 cities %*

Improving Profitability of a Bank – Transcript (1/2)

Interviewer: Your client is a bank facing a reduction in profitability. How would you analyse this?

Candidate: Can you provide more details about the products and services they offer?

The bank offers a range of products like savings/ current accounts, FDs, loans, and credit cards.

Understood. Is there an overall reduction across products, or are any specific products affected? Also, is the reduction in profitability observed throughout the country, or is it a region-specific problem?

There is a significant increase in defaulting loans. The reduction in profitability is observed across the country, but is most pronounced in branches in a specific state.

To understand the problem better, let's break down the profits into revenue and cost. Any significant change in either of them?

The primary issue has been an increase in costs. The costs have risen due to increased provisions for fraudulent bill discounting transactions. Bill discounting is a financial practice where businesses sell their accounts receivable to a third party, typically a financial institution, at a discount. This provides immediate cash to the business.

That's a critical issue. We should investigate the internal processes around bill discounting. Were there any changes in the policies?

There were no significant policy changes. But we did notice some lapses in adherence to procedures.

Process controls should be tightened in parallel, but given the transaction volume, can we use an automated ML model to predict the probability of fraudulent bill discounting rather than relying on manual review alone?

Sure. How should we do that?

Before diving into the model specifications, I'd like to ask a few questions. Could you provide some context about the data we have available and any specific fraud indicators we should consider?

We have a dataset of bill discounting transactions with the following features: transaction amount, client history, transaction date, payment date, previous defaults, unusual payment patterns, and external credit ratings. We also have labelled data indicating whether a transaction was fraudulent.

That is a comprehensive dataset. I'd perform exploratory data analysis to understand the data. Beyond summary statistics and visualisations, I'd specifically check the distribution of transaction amounts, temporal patterns, and the overall class balance. One thing worth flagging is that bill discounting fraud often manifests later, when the bill matures unpaid, so labels can be right-censored. If this is material in the data, a survival-analysis framing may be more appropriate than vanilla classification.

That's a good point, but let's focus only on binary classification for now. What would your next steps be?

Next, I'd preprocess the data by handling missing values, encoding categorical variables, and normalising numerical features. This ensures that the data is in a suitable format for model building. After preprocessing, I would split the data into training and testing sets. I would then look at the class imbalance since fraudulent transactions are likely to be rare compared to legitimate ones.

Good. How would you handle class imbalance?

I'd address class imbalance by using techniques such as

1. **Algorithmic Adjustments –** Applying class weights in the algorithms to penalise misclassifying the minority class more heavily. This is typically the most effective approach for tree-based models and avoids fabricating data.
2. **Resampling –** Either over-sampling the minority class (fraudulent transactions) or under-sampling the majority class (non-fraudulent transactions). Useful when class weights aren't sufficient, but under-sampling discards information, and over-sampling can cause overfitting
3. **Synthetic Data Generation –** Using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic instances of the minority class. More useful for distance-based or linear models. For tree-based models, synthetic interpolation can fabricate implausible fraud patterns

Importantly, resampling and SMOTE should only be applied to the training set after the split, to avoid information leakage into the test data.

That's correct. How would you proceed to build the model?

To build a model for predicting fraud, I'd consider using classification algorithms such as Random Forest and Gradient Boosting Machines. I'd also look at feature importance to understand which features are most indicative of fraud. I'd also use SHAP values to explain individual predictions, which is important both for the risk team's review process and for regulatory audit.

Once you've built your models, how would you select the best one?

I'd use temporal cross-validation, i.e. training on earlier periods and validating on later ones to mirror how the model will be used in production and prevent future information from leaking into

Improving Profitability of a Bank – Transcript (2/2)

the training set. For evaluation metrics, I'd weight PR-AUC most heavily as the ranking metric, since ROC-AUC can be misleading under heavy class imbalance. ROC-AUC can look strong even when the model performs poorly on the rare positive class. I'd also report precision, recall, and F1 score for completeness.

Good approach. How would you deploy this model in a real-world banking environment?

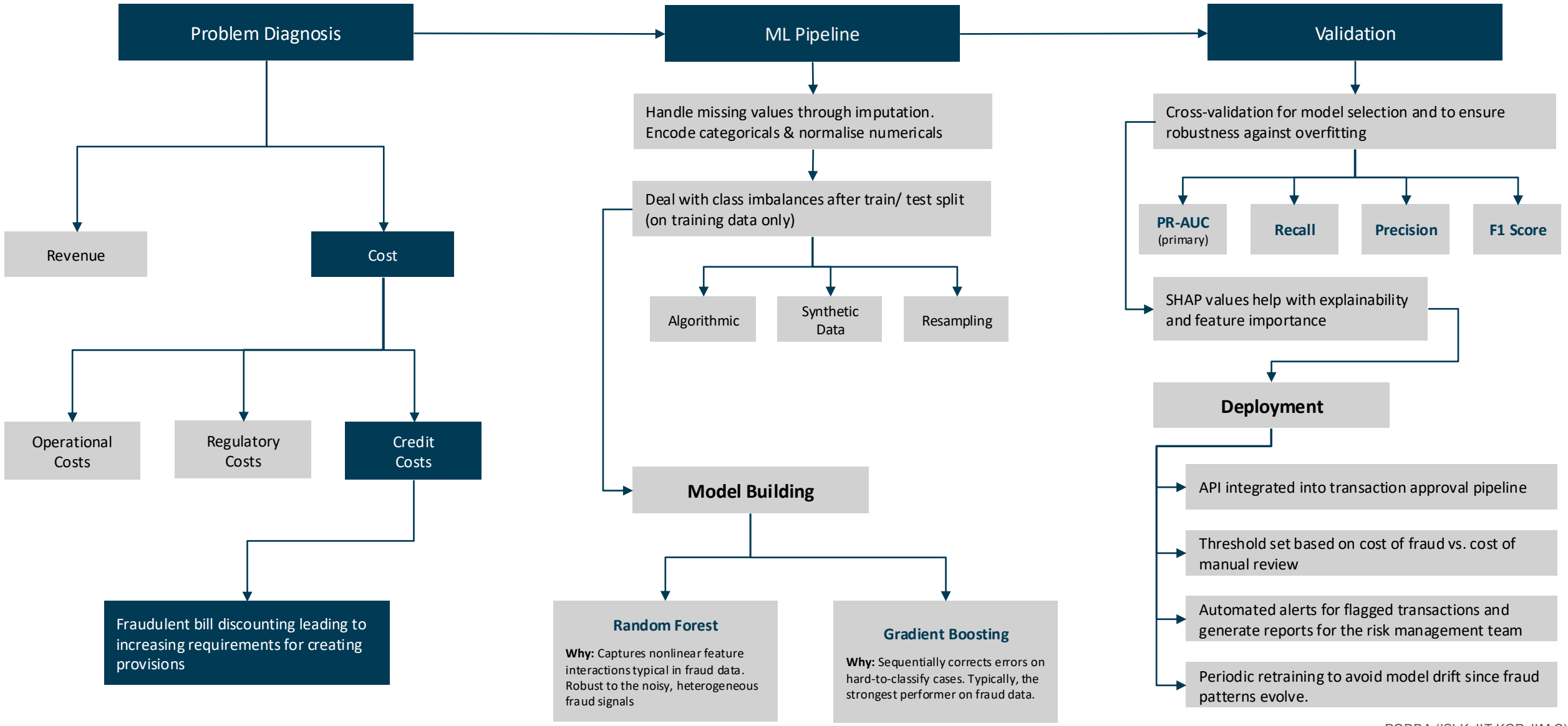
For deployment, I would integrate the model into the bank's transaction processing system. This involves:

1. **Creating an API –** Take transaction data as input and return a fraud probability score.
2. **Setting Thresholds –** Set the threshold by minimising expected cost, weighing the cost of missed fraud against the cost of manual review for false positives.
3. **Alerts and Reports –** Set up automated alerts for flagged transactions and generate regular reports for the risk management team.
4. **Monitoring –** Continuously monitor the model's performance in production and retrain with new data to maintain accuracy.

By reducing fraud-driven loan provisions through automated screening, the bank should see a meaningful improvement in cost structure and therefore profitability, quantifiable as the expected fraud loss avoided net of model operating cost.

Good suggestions. Thank you.

Improving Profitability of a Bank – Case Flow



Mitigating Credit Card Churn – Transcript (1/2)

Interviewer: Your client is a financial technology company offering corporate credit cards. They've seen a recent increase in customer churn, especially among startups. How would you approach identifying strategic measures to mitigate these challenges?

Candidate: Before diving in, I have two clarifying questions: what's the client's value proposition and key differentiator, and how is churn currently defined?

The client offers a seamless financial experience tailored to startups, with higher credit limits without personal guarantees, simplified expense management, and rewards for fast-growing companies. Churn is defined as account closure or 90+ days of zero transaction activity, whichever comes first.

What's changed recently in the competitive landscape, and which customer segments are most affected?

There's increased competition from newer fintechs offering similar benefits at lower fees. Early-stage tech and e-commerce startups are most prone to churning, particularly when facing cashflow challenges or slower-than-expected growth.

Are there underused features or recurring feedback themes? And has any of this been systematically tied back to churn behaviour?

Expense management tools and financial analytics are underused. Customers often mention wanting more tailored financial advice and quicker issue resolution. Feedback is collected, but it hasn't been systematically tied to churn.

Are there any economic factors that might be disproportionately affecting certain industries within the customer base?

Some industries have indeed been hit harder by recent economic downturns, which might influence their financial behaviours and contribute to churn.

That’s important to consider. Once we’ve identified these root causes, we can move on to predictive modelling to identify at-risk customers. What kind of data does the client currently collect that we could use to build such a model?

They collect transaction history, spending patterns, reward redemption, customer service interactions, demographic details, and industry-specific information.

That’s a comprehensive dataset. Before modelling, I'd run an exploratory analysis to understand the data. I would look at the overall churn rate by segment, the distribution of customer tenure, and behavioural patterns preceding churn, like declining transaction frequency, lower spend, fewer logins, and increased support tickets. I'd also check the time-to-churn distribution, since this informs whether we frame the problem as classification or survival analysis.

How would you choose between the two?

1. **Classification –** Predicting churn within a fixed window, typically 60 - 120 days, based on business response cycle. It is simpler to deploy and integrates cleanly with retention workflows
2. **Survival Analysis –** Gives a richer view of when a customer is likely to churn, which helps prioritise outreach

I'd start with classification for simplicity and add survival later if the timing of intervention matters.

Let's focus only on classification for now. What important factors will you consider before building the model

1. **Class imbalance –** churn rates are typically 5 - 15%, which is not severe. I'd start with class weights before experimenting with resampling. SMOTE-style synthetic interpolation can fabricate implausible churn patterns
2. **Temporal validation –** train on a past window and validate on a future window, not a random split, to mimic production
3. **No data leakage –** engineer features only from data available before the prediction date, to avoid look-ahead bias

Sounds good. How would you build the classification model?

Since the client hasn't used models before, I'd start with logistic regression as a baseline for interpretability and a sense check on feature directions. I'd then move to XGBoost, which typically outperforms on tabular churn data, and use SHAP for feature-level explanations that match logistic interpretability without sacrificing accuracy.

How will you select the right model?

I'd weight recall and precision based on retention-offer economics. I’d favour recall when offers are cheap and precision when they're expensive. PR-AUC is the primary summary metric, with precision, recall, and F1 score reported alongside for completeness.

Beyond churn probability, I'd also score each customer by expected CLV at risk = P(churn) × CLV, so retention spend prioritises high-value accounts rather than just high-probability churners. Regular retraining and A/B testing of retention interventions can ensure the model remains effective as customer behaviour changes.

Mitigating Credit Card Churn – Transcript (2/2)

Before we move to the retention strategy, you mentioned survival analysis as an alternative framing. Please elaborate on that.

Survival analysis would shift the question from "will this customer churn in the next 90 days?" to "how long until this customer is likely to churn?" I'd use a Cox proportional-hazards model, which estimates a hazard rate over time as a function of customer features. It handles right-censoring naturally, and the time-to-event output lets us sequence interventions. We can use cheap automated nudges early in the hazard curve, and more expensive personalised offers as the hazard rises. The trade-off is added complexity in deployment, since the output is a curve rather than a single score.

Good. Now, let’s move on to the retention strategy.

I'd recommend four retention strategies:

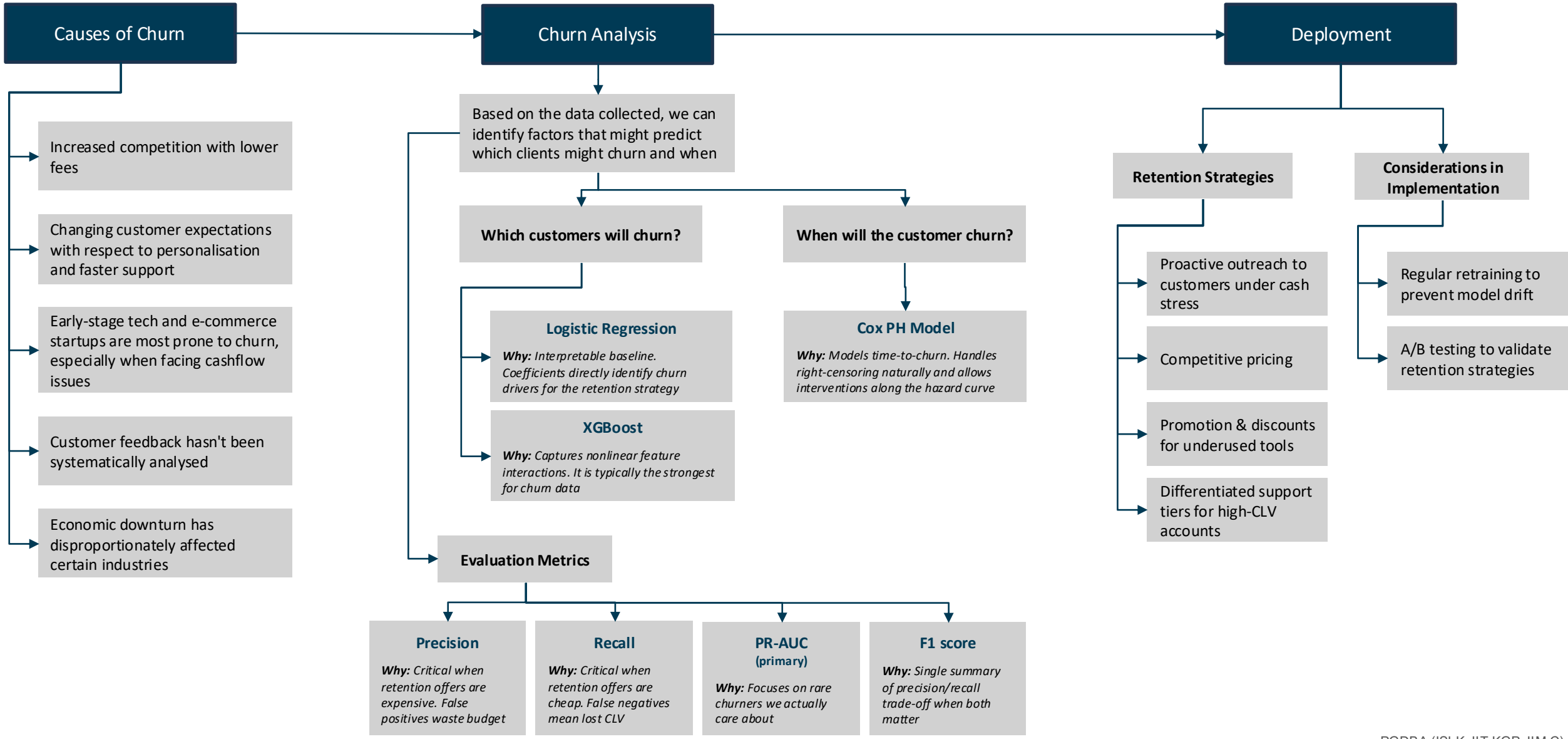
1. **Proactive outreach** to early-stage startups under cashflow stress
2. **Fee adjustments** where competitors are winning on price
3. **Promotion and discounts** for underused tools
4. **Differentiated support tiers** for high-CLV accounts

How would we know which of these retention strategies is effective?

They should be A/B tested, with retained CLV net of cost as the success metric.

Good recommendations. Thank You. We can close the case here.

Mitigating Credit Card Churn – Case Flow



Optimising Pharma Supply Chain – Transcript (1/2)

Interviewer: Your client is a pharmaceutical company. They are unable to meet their current demand. What would you suggest they do?

Candidate: I would like to know some information related to the kind of products they manufacture.

They manufacture and distribute a wide range of prescription medications, with a particular focus on chronic disease treatments such as diabetes and hypertension.

Ok. The value chain of a pharmaceutical company would be: product development, procurement of raw materials, manufacturing, storage, distribution, marketing, and sales. Can you tell me where in the value chain they are facing a problem?

Recently, the company has been facing challenges with stock-outs at various distribution centers, leading to significant delays in delivering products to pharmacies and hospitals. Can you identify the root cause of these stock-outs and suggest ways to optimize their supply chain and distribution processes?

So, I need to identify why the stock-out rates are high and propose strategies to improve supply chain efficiency. Can you tell me how long this has been an issue?

Yes, that’s correct. The problem has been worsening over the past two years.

Understood. To get a clearer picture, could you provide more details about your distribution process, your main customers, and whether there are specific regions or products where the stock-outs are more frequent?

Our distribution network covers the entire country, with 12 regional distribution centers. Our customers include large pharmacy chains, hospitals, and some independent pharmacies. We’ve noticed that stock-outs are more frequent in rural areas and for certain high-demand medications like insulin.

Thanks for the details. Before we proceed, I want to confirm: How are stock-outs defined? Is it any instance where an order cannot be fulfilled within a certain timeframe?

A stock-out occurs when a product is unavailable for delivery within 48 hours of order placement

I would examine the entire distribution network, starting with demand forecasting and inventory management at the distribution centres, and then consider potential bottlenecks in transportation or supplier reliability. I’d also look into the replenishment process to understand where exactly the problem lies.

That sounds like a good plan. We've already investigated supplier and transport bottlenecks and ruled them out. We’ve also looked into inventory management and the replenishment process. The remaining issue is forecast accuracy, particularly for high-criticality SKUs at rural DCs. Let's restrict our analysis to demand forecasting. How should we proceed?

Before jumping to a model, I'd start with EDA on the historical demand data. Specifically, I'd look at demand trends and seasonality at the SKU and distribution-centre level, the volatility of demand for high-stock-out products like insulin, the correlation between stock-out events and external factors, and any structural breaks in the series. This shapes which forecasting approach is appropriate.

That makes sense. What forecasting methods would you consider?

I'd start with SARIMAX as a baseline, as it captures seasonality, which is relevant for chronic-disease and respiratory medications, and accommodates exogenous regressors like prescription trends, promotions, and epidemic indicators. SARIMAX is interpretable, stable on series with regular demand, and a sensible starting point given the client hasn't used statistical forecasting before.

Is SARIMAX the best model here?

SARIMAX assumes a continuous, regularly recurring demand pattern. Many pharma SKUs, particularly slow-movers and specialised treatments, exhibit intermittent demand, with frequent zero-demand periods between orders. SARIMAX is unstable on these series. For that segment, I'd use Croston's method, which separately models the demand size and interval between non-zero orders. The choice between SARIMAX and Croston's would be driven by the demand pattern at the SKU level

That’s a good point. Now could you suggest some metrics for evaluation?

We can use MAE, as it is a straightforward way to measure the average error in our forecasts. We can use MAPE to compare forecast accuracy across different products, since it is scale-free.

And how would you validate the model itself?

I'd use rolling-origin cross-validation rather than a single train-test split, since demand forecasting is fundamentally time-dependent and a random split would leak future information into training. The evaluation window should reflect how the forecast is used in production, so I'd validate at the same horizon the company plans for replenishment, typically two to four weeks.

I'd also asymmetrically weight errors because under-forecasting is far more costly than over-forecasting. I’d use pinball loss at the 90th percentile, since it penalises under-forecasting roughly 9×

Optimising Pharma Supply Chain – Transcript (2/2)

more than over-forecasting, which matches the asymmetric cost of stock-outs vs. holding excess inventory.

Those metrics make sense. How do you think demand forecasting can help with our problem?

It would help in reducing stock-outs, and better forecasts can help lower holding costs and minimise waste, especially for perishable items. It would also increase customer satisfaction by ensuring timely deliveries of medications. Inventory policy itself, such as reorder points and safety-stock rules, is a parallel lever worth reviewing, but forecasting is typically the larger contributor to stock-out rates.

Absolutely. How would you suggest we begin implementing these changes?

I recommend starting with a pilot project for the client's most critical products, like insulin. We can develop and test a forecasting model using historical data and relevant external factors, and if successful, scale it across the entire product range and distribution network.

Good. How would we measure whether the pilot succeeded?

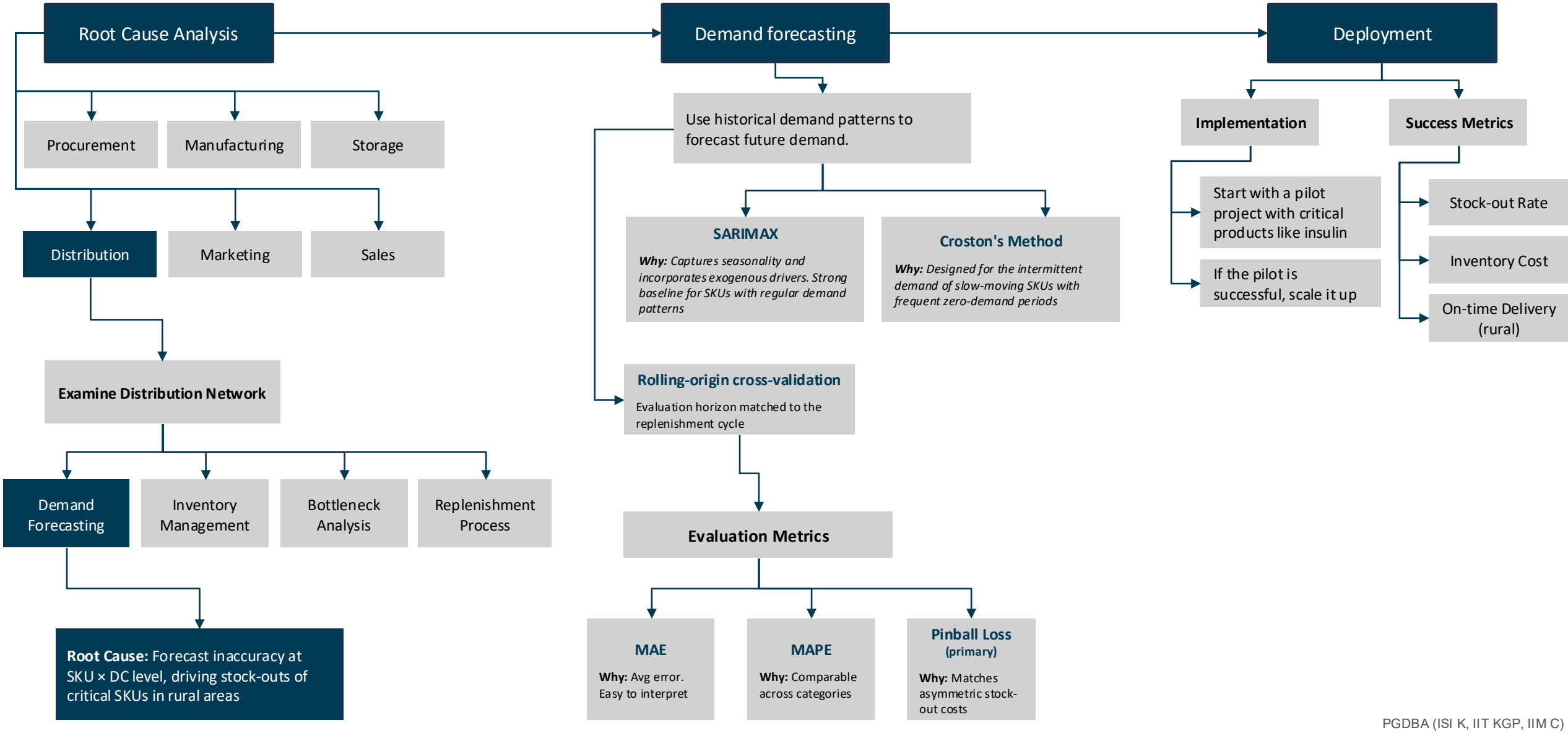
The success metric should map directly to the business problem, not just forecast accuracy. I'd track three things.

1. **Reduction in stock-out rate** at piloted distribution centres, particularly for the critical SKUs
2. **Inventory holding cost** since over-forecasting has its own cost in pharma
3. **On-time delivery rate to rural areas**, where the problem is most acute

If the pilot delivers a meaningful improvement on stock-out rates without inflating holding costs, it's worth scaling. Forecast accuracy (MAPE, pinball loss) is a leading indicator, but stock-out rate and service level are what the business actually cares about.

That's a solid plan. We'll consider starting with a pilot project as you suggested. Thank you for your insights.

Optimising Pharma Supply Chain – Case Flow



Reducing Heart Failure Readmissions – Transcript

Interviewer: You're a data scientist at a large hospital. The hospital has observed an increase in readmission rates within 30 days of discharge for patients with heart failure. The administration wants to understand the reasons behind this increase and identify potential solutions. How would you approach this problem?

Candidate: Before diving into the analysis, I'd like to clarify a few points. May I ask some questions about the available data and the current situation?

Of course, please go ahead.

Over what time period has this increase in readmission rates been observed? Do we have any initial hypotheses about potential causes?

We've noticed this trend over the past 12 months. As for hypotheses, there are concerns about the quality of discharge planning and post-discharge support, but we're open to your analysis.

To start, I would do an EDA to understand the distribution and trends in our dataset. This would include summarising key statistics, identifying missing values, and looking at correlations between different variables. Specifically, I'd look at readmission rates over time, distribution of length-of-stay, missingness patterns in follow-up care data, and the class balance for the readmission outcome.

That's a good start. How would you proceed next?

I would like to segment the data into different cohorts, such as age groups, severity of heart failure, comorbidities, and socio-economic status. This will help us identify if certain groups are more prone to readmission. Does that make sense?

Yes. How would you determine the key factors contributing to the increased readmission rates?

I'd benchmark against validated risk scores like LACE or HOSPITAL first. Any custom model needs to outperform them. For the custom build, I'd start with logistic regression as a baseline, then XGBoost with SHAP for stronger performance with interpretability. Both are predictive, not causal, and so I'd layer in propensity score matching to estimate the effect of discharge planning quality, or difference-in-differences if a policy change is available.

How would you validate your findings?

I'd split the data into training and testing sets and use cross-validation to ensure robustness. I'd use temporal validation rather than a random split since care protocols and patient mix evolve over time, and a random split would leak future information. The AUC- ROC metric would be ideal here, as it evaluates the model's ability to balance sensitivity and specificity across different thresholds. The AUC-PRC metric might be more useful if the data is heavily imbalanced. I'd also check calibration, since clinical decisions depend on absolute probabilities rather than rankings.

Once the key factors are identified, what recommendations would you provide to the hospital?

Based on the analysis, I'd recommend targeted interventions for identified high-risk groups. For example, improving follow-up care for elderly patients or providing additional resources for those with severe heart failure. Implementing a pilot program to test these interventions before full-scale deployment will be necessary.

How would you design the pilot program for the interventions?

The pilot would use a cluster-randomised design, randomising at the ward or care-team level rather than the patient level, to avoid withholding a plausibly beneficial intervention from individual high-risk patients and to prevent contamination from staff trained on the new protocol. We'd monitor readmission rates, patient satisfaction, and health outcomes to measure effectiveness.

What challenges might you anticipate during this project, and how would you address them?

Challenges like data quality and patient compliance are expected. We'll address data issues with thorough cleaning and appropriate imputation techniques. For compliance, interventions must be user-friendly. Additionally, engaging and training medical staff will be crucial to overcoming any resistance

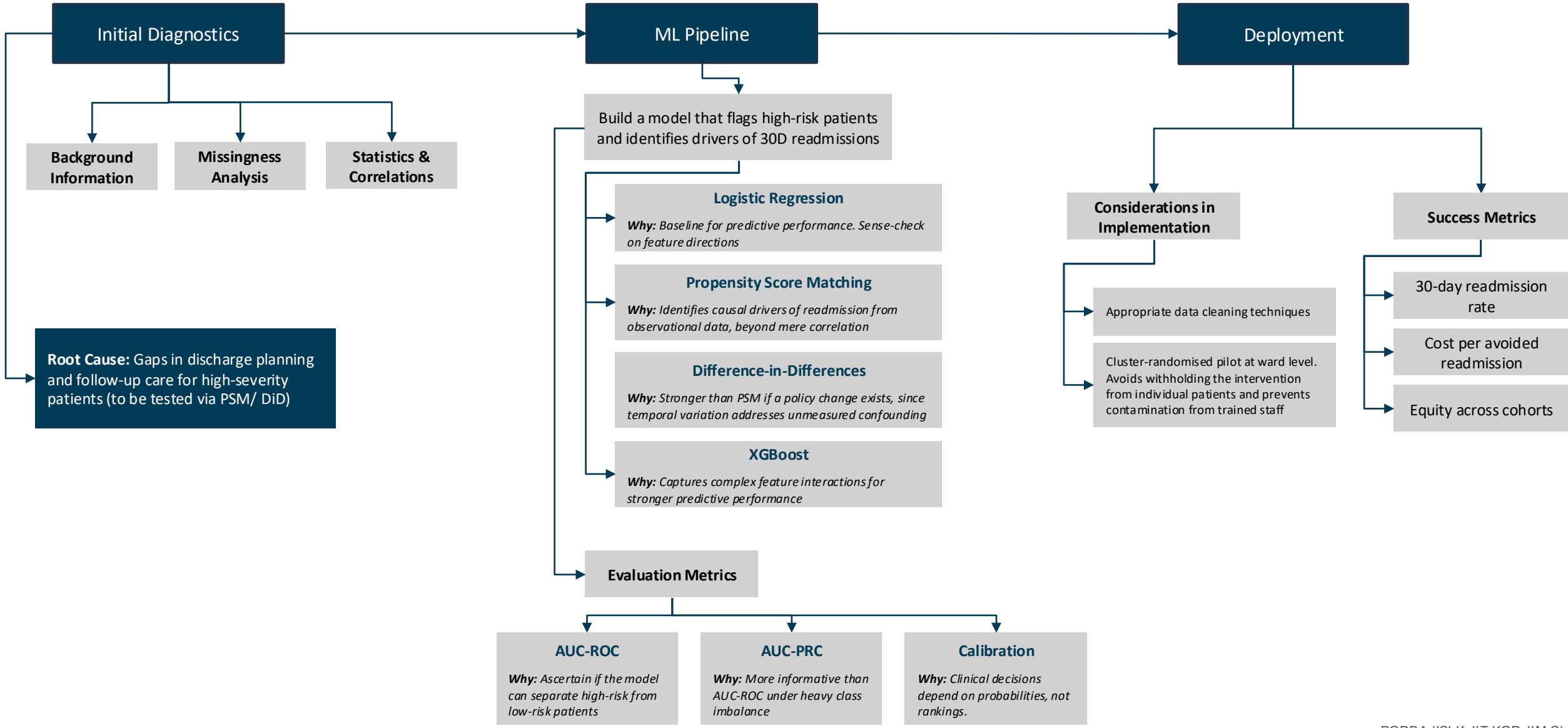
How would we measure whether the programme is succeeding?

I'd track three things:

1. **30-day readmission rate** for the heart failure cohort, particularly in the high-risk segments
2. **Cost per avoided readmission**, since interventions like enhanced follow-up care have their own cost and the programme has to deliver net savings
3. **Equity across cohorts**, ensuring readmission reductions are realised across socio-economic groups rather than concentrated in those easiest to reach

Thank you. That concludes our case discussion.

Reducing Heart Failure Readmissions – Case Flow



Retaining Social Media App Users – Transcript (1/2)

Interviewer: Your client is an upcoming Indian regional social media app. They want to explore revenue streams.

Candidate: I'd like to ask a few questions before we get into it. Could you specify which regions and languages the app is focused on? What are the key features of the app, and what are the current revenue streams?

The app is limited to South India, specifically Malayalam, Tamil, Kannada, and Telugu. It features only short-form videos. The app is mostly used by young adults aged 18-30. The current revenue comes from in-app purchases.

Got it. I'd also like to understand the current user engagement metrics such as daily active users (DAU), monthly active users (MAU), average session duration, and retention rates. Additionally, do we have any insights into how our competitors are performing?

We have basic engagement metrics but no detailed competitor analysis yet.

Understood. I think the major source of revenue, like any social media app, would be from advertising.

Yes, I want you to delve deeper into how we could pitch our app to advertisers. Focus on who our potential advertisers would be and the value proposition.

Given our regional focus, we can target two sets of brands. First, we can target regional brands with a strong presence in South India and national brands looking to penetrate this market. Our value proposition to advertisers could include:

1. **Breaking the language barrier –** Access to a highly engaged audience that is often untapped by other platforms due to language limitations
2. **Young user base –** Primary demographic is users aged 18-30, which is attractive for brands aiming for long-term value
3. **Engaging content format –** the short-form video format aligns with modern users' limited attention spans, making it ideal for effective ad placements.

The client is also facing customer retention issues. Strong ad revenue depends on a sticky user base, so let's look at this next. How can they improve the time spent by users on the app?

Before jumping to solutions, I'd run some exploratory analysis on the engagement data. Specifically, I'd look at the retention curve by cohort (D1, D7, D30 retention by language), session duration distributions, critical points in the user journey (signup, first video, first follow), and whether engagement varies systematically by language. This shapes which retention lever to prioritise.

Sounds good. Once you have those findings, what would you focus on?

The main causes of low engagement I'd hypothesise are poor content relevance, language mismatch, cold-start friction, dormant users, app quality issues, and a thin social graph.

Let's focus on poor content relevance and dormant users.

For poor content relevance, we need to improve relevance ranking. I'd combine two models: a watch-time regression as the predictive base layer, and a lightweight RL layer on top to optimise for return visits rather than just immediate watch time.

Why watch-time regression, and how would you implement it?

Watch time is a much better proxy for "time on app" than click-through rate, because clicks alone reward clickbait. A user can click a thumbnail, watch for 3 seconds, and move on. Watch time captures whether the content actually held attention.

Implementation-wise, I'd train a logistic head on engaged vs. non-engaged events, with positive examples weighted by watch time. The model outputs odds, which under low click-probability conditions approximate expected watch time per (user, video) pair. This is essentially the same technique that large social media platforms use and gives us a strong baseline that's easy to ship and debug.

And why add an RL layer on top?

The watch-time regression has a known failure mode. It'll over-rank content that maximises immediate watch time, which often means clickbait clips that hold attention in the moment but don't bring users back. The RL layer reframes the feed as a sequence of decisions, i.e. each recommendation is an action, and the policy is rewarded when users return the next day, not just when they watch a long video now.

I'd start with something lightweight like contextual bandits over candidate videos already shortlisted by the regression model, with the reward being a multi-day return signal rather than session-level watch time. This guards against the regression model amplifying clickbait while keeping the system tractable. A full deep RL would be over-engineering for this stage.

Interesting. Now let's focus on churn prediction.

Retaining Social Media App Users – Transcript (2/2)

For churn prediction, I would use a gradient-boosted classifier like XGBoost or LightGBM trained on tabular engagement features: friends, followers, time spent on app, notification click-through, days since last login, watch-time trend, language-activity mix, etc. Boosted trees are preferable to deep models here for interpretability, as SHAP values let the growth team see why a user is at risk, and for sample efficiency on a modest user base. High-risk users are then routed to incentives such as free rewards, re-engagement pushes, or notifications timed to their historical active hours.

How would you validate these models before rolling them out?

Validation has to be temporal. I'd track MAE for the watch-time regression, and for the churn classifier, I'd use AUC for overall ranking quality alongside precision@k since the growth team can only act on a fixed budget of users, with calibration checks so predicted probabilities match observed rates.

For the RL policy, I'd use off-policy evaluation via inverse propensity scoring, re-weighting historical data to estimate how the new policy would perform without exposing users to it, as an offline check before any live A/B test.

Interesting approach. What else can be done? And name some success metrics that we should look for.

Beyond modelling, we should collect direct feedback via Play Store reviews and in-app surveys, run sentiment analysis to flag issues no model can fix, like bugs and missing features, and validate every change with A/B tests.

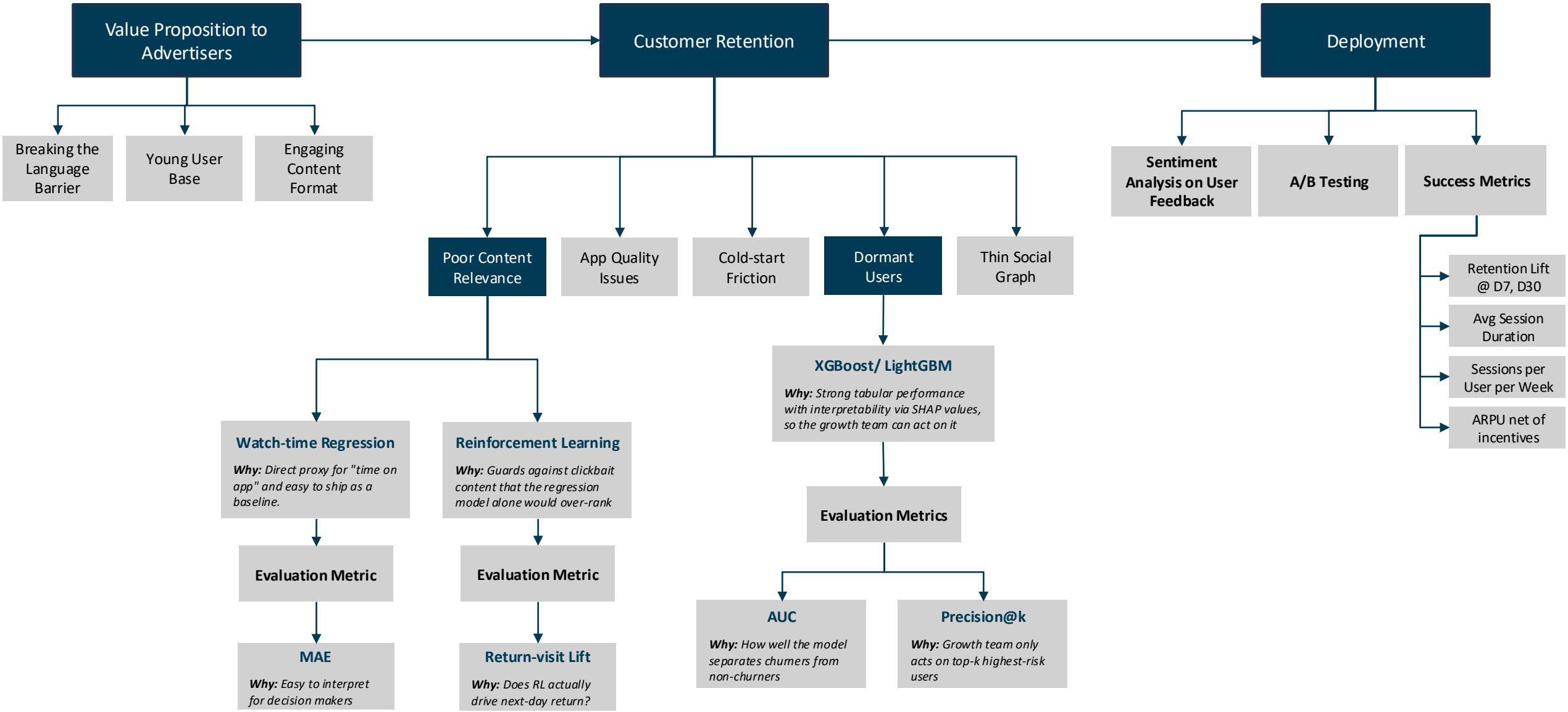
And how would we know whether the retention programme is actually working?

Success should map back to the business problem, not just model accuracy. I'd track three things:

1. **Retention rate lift at D7 and D30** for treated cohorts versus control, since these are the standard milestones for habit formation
2. **Average session duration** captures the relevance and engaging nature of recommendations
3. **Sessions per user per week** capture the user's satisfaction. Users who are not satisfied will not revisit
4. **ARPU change net of incentive cost**, since free awards and re-engagement pushes have a real cost and the programme has to deliver positive net revenue. Model metrics like AUC and MAE are leading indicators, but retention lift and net ARPU are what the business actually cares about.

Thank you, that will be all.

Retaining Social Media App Users – Case Flow



Retaining Mobile Gaming Customers – Transcript (1/2)

Interviewer: Your client is a mobile gaming company, and they've been struggling with player churn, especially over the past six months. How would you approach this case?

Candidate: I'd start by confirming the scope and magnitude of the churn issue. Is it affecting all player segments or just specific ones, like new players, medium-term players, or long-term players? Additionally, I'd gather details on when this churn is happening, whether it's during onboarding, after a few days of gameplay, or much later.

That's a good starting point. Let's assume churn is predominantly in the first 30 days.

That's helpful context. If most churn is happening within the first 30 days, my focus would shift to understanding player engagement during this critical period. I'd analyse metrics like D1, D7, and D30 retention rates, which would highlight when exactly players are dropping off. I'd also want to look at session frequency and average session length to understand if players are becoming disengaged early on. Are there specific levels or in-game moments where we're seeing these drop-offs?

Yes, there seems to be a steep decline in engagement after the first few gaming sessions, especially in the tutorial and early stages. What would you suggest based on this?

Based on that, I'd suggest two initial areas to investigate. First, the onboarding process. It's likely that players are either not finding the game mechanics engaging or they're feeling overwhelmed early on. Testing different onboarding experiences through A/B testing could provide useful insights. We could optimise tutorials, simplify early gameplay, or add early-stage rewards to hook players. Second, I'd assess the game's difficulty curve. Smoothing it out by introducing easier challenges in the beginning and gradually increasing complexity could help retain more players.

Interesting. Let's say the client agrees to improve onboarding. They're also worried about losing their high-value players. How would you address these concerns?

High-value or "whale" players are crucial, so for them, I'd recommend a more personalised approach. Offer recommender trained on each whale's past purchases and engagement, so VIP content is personalised per player rather than one-size-fits-all. Anomaly detection on rolling spend or sessions using isolation forest or z-score to catch early disengagement before it shows in aggregate metrics. If we detect early signs of disengagement, like fewer sessions or reduced spending, immediate interventions like push notifications or personalised offers could be triggered to re-engage them.

Good approach. How would you measure the effectiveness of these initiatives?

I would track changes in D1, D7, and D30 retention rates for new cohorts. For whales, I'd focus on changes in Lifetime Value (LTV) and Average Revenue Per User (ARPU). I would compare the Churn Rate and Average Session Duration before and after implementing changes. We could also use Cohort Analysis to segment players and assess improvements in retention and engagement among both new players and high-value players.

That makes sense. Would you also consider predictive models in this scenario and what kind of features would you like to include?

We could develop a churn prediction model using features like session frequency, in-game purchases, and progression levels. XGBoost or LightGBM works well, chosen for interpretability via SHAP values, so the live-ops team understands why each player is flagged.

For whales, I'd train a separate survival model, such as a Cox proportional hazards model or gradient-boosted survival trees, which treats time-to-churn as a continuous variable and outputs a survival curve. It also handles right-censoring correctly, so whales still active at the observation cutoff are properly accounted for. Predicting when a whale will churn is more useful than a binary flag, as it lets the team intervene days before the predicted churn point.

How would you validate these models before rolling them out?

Validation has to be temporal, since player behaviour and game content shift with each update. I'd train on earlier cohorts and validate on more recent ones. For the churn classifier, I'd track AUC and precision@k, since the live-ops team only acts on the top-k highest-risk players, and check calibration so predicted churn probabilities match actual rates.

For the survival model on whales, I'd use time-dependent AUC at the relevant intervention horizons (e.g., 7 and 14 days), since live-ops acts within a specific window. C-index gives a single-number summary across horizons as a secondary check. Finally, I'd evaluate fairness and stability across player segments to make sure the model isn't systematically worse for any one group.

How would you ensure that the model's predictions translate into actionable insights?

Once the model identifies at-risk players, we'd trigger automated retention strategies like personalised offers or in-game rewards via push notifications. The intervention itself can be optimised with a contextual bandit, given a player's state (segment, recent activity, spend tier), the policy chooses which offer type to surface and learns over time which intervention works best for which segment, rather than relying on a single fixed playbook. I'd retrain weekly and monitor for feature drift, since player behaviour shifts with each game update.

Retaining Mobile Gaming Customers – Transcript (2/2)

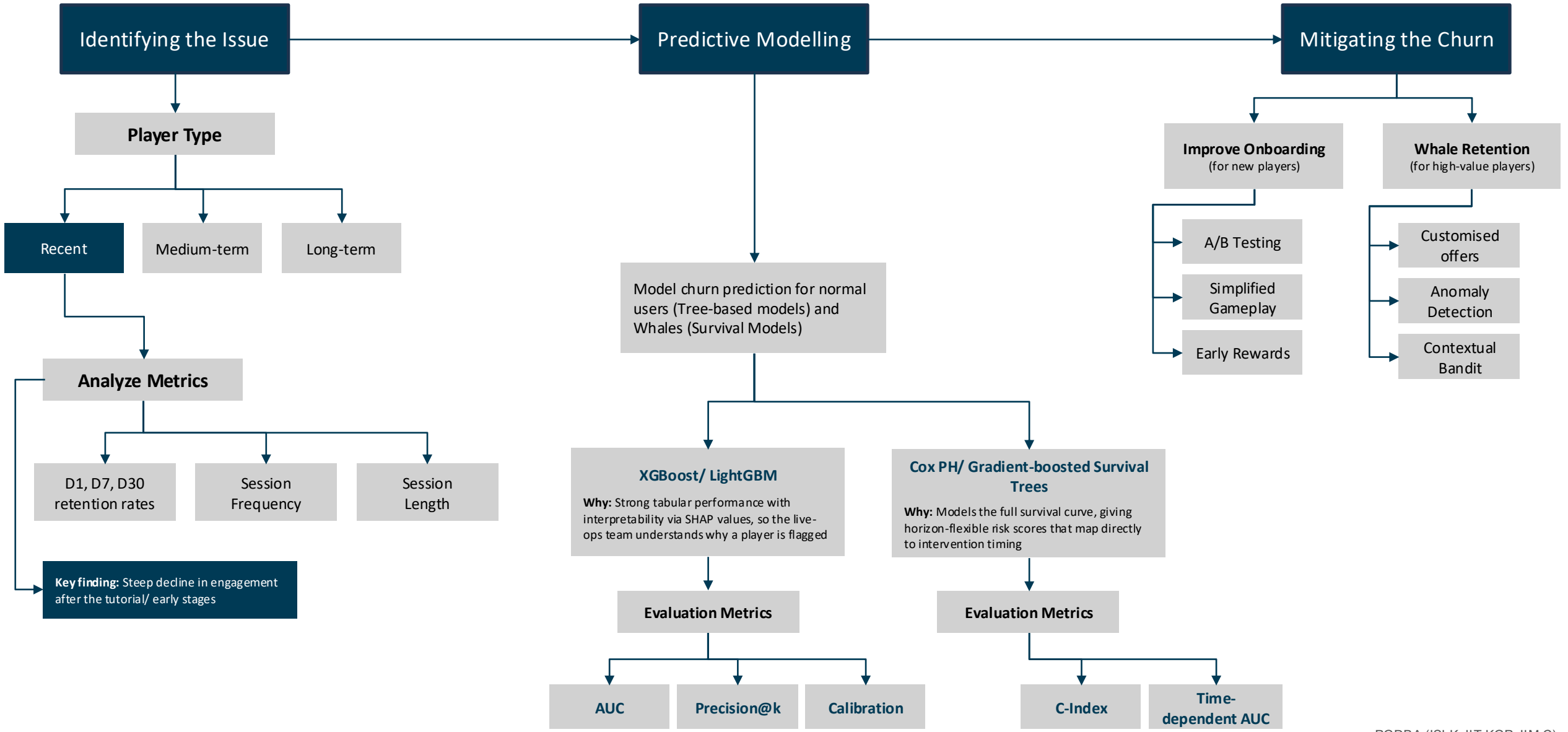
That makes sense. What's the key takeaway for the client here?

The retention programme should be measured against three things:

1. **D7 and D30 retention lift** for new cohorts versus control
2. **LTV and ARPU change for whales net of intervention cost**, since personalised offers and VIP content have a real cost, and the programme has to deliver positive net revenue
3. **Churn rate stability across player segments**, ensuring gains aren't concentrated in one geography or platform

Excellent, thank you. That was very insightful.

Retaining Mobile Gaming Customers – Case Flow



Sales Forecasting for an E-comm Platform – Transcript

Interviewer: Today, we're going to dive into some data challenges that your client's e-commerce platform faces. They specialise in lifestyle and home products, and they've gathered quite a bit of data over the years. How would you approach cleaning and preparing this data for analysis, given that it includes transactional data, product details, customer info, and web traffic statistics?

Candidate: To start, I'd merge the transactional data with product and customer details to create a comprehensive dataset. Handling missing values would be crucial. I'd look at the importance of each field and decide whether to impute or remove these values. Normalising prices across different markets and categorising products could provide clearer insights for comparison.

Sounds like a solid plan. Let's talk about exploring this data. What specific metrics and visualisations would you use to understand sales trends and customer behaviours?

I'd focus on year-over-year sales growth and categorise sales by product types. Visualising this with bar charts and line graphs would highlight trends and outliers. Additionally, looking at customer purchase patterns and segmenting by demographics could reveal key insights into the target market's buying habits, especially during peak seasons.

I like where this is going. Given your findings, how would you predict sales for the upcoming holiday season? What modelling techniques and variables would you consider?

For forecasting, I'd suggest a SARIMAX model to handle the seasonality of the sales data. Exogenous variables would include historical sales, price points, promotional data, and categorical data like product types. As a benchmark, I'd also test a Random Forest with engineered lag and rolling-window features to capture nonlinear interactions that SARIMAX might miss.

I'd validate using walk-forward cross-validation, with the horizon matched to the replenishment cycle, and track WAPE alongside a pinball loss at the 90th percentile because under-forecasting during the holiday season costs more than over-forecasting, so the loss should reflect that asymmetry.

That's very thoughtful. Shifting gears to inventory management, how would you suggest the client use their data to optimise their inventory for high demand during the holidays?

I would analyse historical sales to predict future demands, focusing on high-turnover products. It's important to calculate optimal stock levels, especially considering potential supply chain delays. For products with unpredictable sales, maintaining a safety stock $\mathbf{z} \times \mathbf{std\ dev} \times \sqrt{\mathbf{lead\ time}}$ might help avoid stockouts, where z is the service-level z-score, σ is the demand standard deviation, and the square root scales for variance accumulating over the lead time

Absolutely, planning ahead is key. Finally, considering marketing, how would you go about segmenting the customer base for more targeted campaigns?

I'd utilise RFM analysis, segmenting customers by Recency, Frequency, and Monetary value. High-RFM customers get retention campaigns, lapsed high-spenders get win-back offers, and low-frequency users get nurture flows. For engagement, targeted email marketing tailored to these segments could be effective. Also, experimenting with different website layouts through A/B testing could optimise conversions across different customer profiles.

Those strategies sound like they could really enhance our customer engagement. Based on your analysis today, what top three strategic recommendations would you give the client to boost sales for the upcoming holiday season?

First, I'd focus on optimising promotions for high-margin products. Second, enhancing the user experience on mobile devices could capture more sales. Finally, increasing marketing efforts in our top-performing regions could leverage local popularity to boost overall sales.

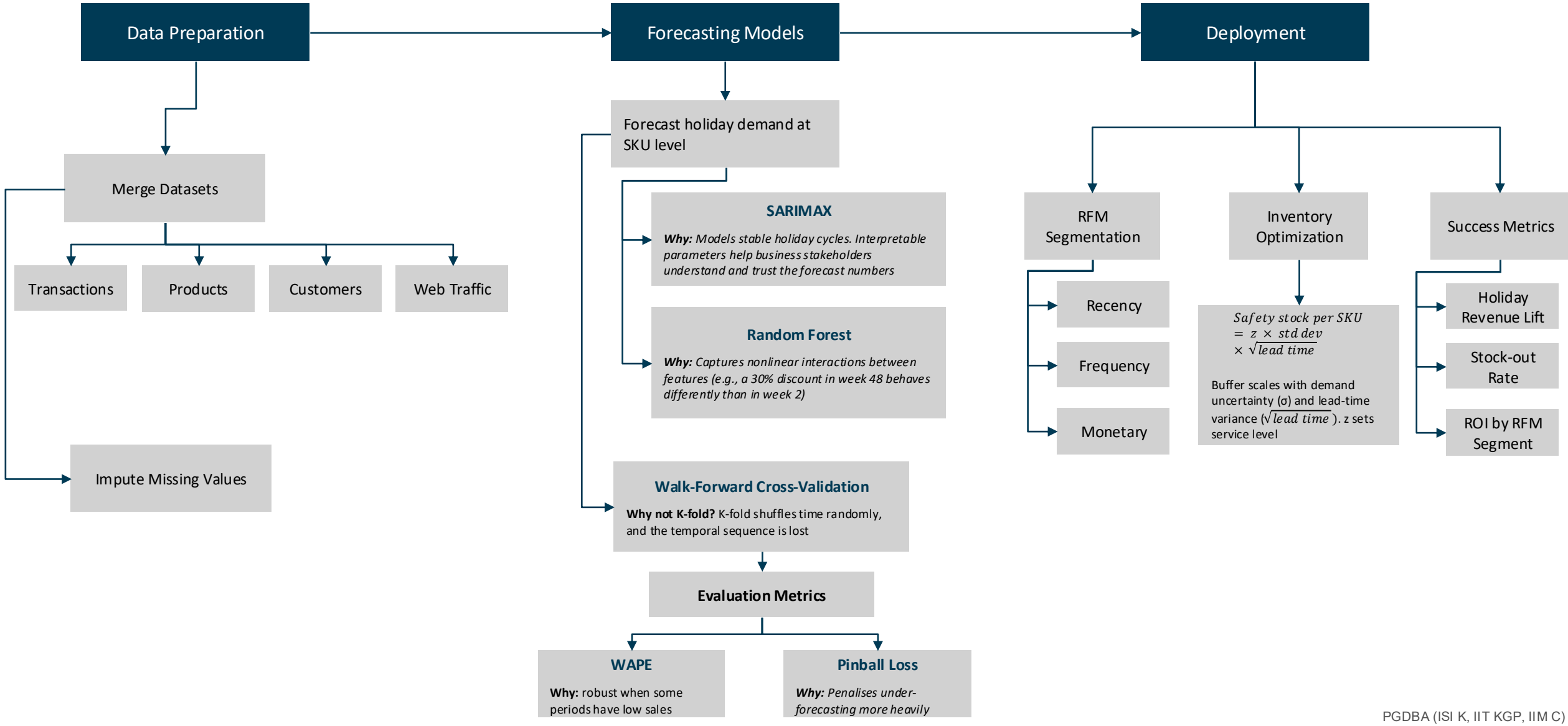
And how would we measure whether these recommendations actually deliver?

I'd evaluate on three success metrics:

1. **Holiday-season revenue lift** versus the same period last year, controlling for traffic growth
2. **Stockout rate and inventory holding cost** for high-turnover SKUs
3. **Campaign ROI by RFM segment**, since retention spend on high-value customers should deliver a higher return than broad-based discounting

Fantastic! Your insights are incredibly thorough. Thank you for such a detailed discussion today.

Sales Forecasting for an E-comm Platform – Case Flow



Reducing Telecom Customer Churn – Transcript

Interviewer: Let's discuss a common challenge in the telecom industry regarding customer churn. Can you explain what customer churn is and why it's important to telecom companies?

Candidate: Customer churn refers to the rate at which customers stop doing business with an entity. In the context of telecom, it's a metric that measures the percentage of subscribers who discontinue their subscriptions within a given time period. It's crucial because acquiring new customers is generally more expensive than retaining existing ones, and high churn rates can significantly affect a company's revenue and profitability.

Excellent explanation. Let's dive into some data analysis. You're given a dataset containing customer demographics, service usage patterns, billing information, and churn status. How would you start analysing this data to understand the churn patterns?

First, I would perform an EDA to understand the distributions of features and their relationship with churn. This would involve summarising statistics and visualising data trends, such as examining how different age groups, billing methods, or usage patterns correlate with churn. Tools like correlation matrices and feature importance plots can help identify which factors are most predictive of churn.

Interesting approach. Suppose your initial analysis shows that customers on month-to-month contracts are churning at a higher rate compared to those on one-year or two-year contracts. What could be your next step to delve deeper into this insight?

To build on that insight, I would segment the customer base by contract type and further analyse the characteristics of customers within the month-to-month group to identify specific risk factors or dissatisfaction points leading to churn.

Additionally, I would apply survival analysis to model time-to-churn directly. Specifically, I'd plot Kaplan-Meier curves stratified by contract type to visualise how survival probabilities diverge over tenure. I'd then fit a Cox proportional hazards model with covariates including contract type, monthly charges, service bundle, payment method, and tenure-at-observation, to quantify the hazard ratio associated with each factor while controlling for the others.

Why not just run a logistic regression?

Logistic regression treats churn as binary within a fixed window, dropping or mislabelling censored observations and collapsing timing, so month 2 and month 22 churners look identical. Survival models handle right-censoring and preserve the time dimension, telling us not just who will churn but when.

How do you propose we address this high churn rate among month-to-month customers?

I would recommend several strategies:

- **Service Quality Improvements –** Address the operational drivers of dissatisfaction surfaced in the EDA, such as network reliability, billing dispute resolution times, or customer-service wait times
- **Incentive Programs –** Offer incentives for month-to-month customers to switch to longer-term contracts, such as discounted rates or enhanced service features
- **Customer Engagement –** Increase customer engagement through regular feedback loops and personalised communications, aimed at addressing their needs and improving satisfaction

Those are thoughtful suggestions. How would you perform predictive modelling?

I would develop a predictive model to identify at-risk customers and target them with specific retention strategies based on their preferences and usage patterns. I would complement this with survival models such as Cox regression or random survival forests, which output an individualised hazard curve per customer rather than a single churn probability. This allows the retention team to prioritise not only which customers to target but also when to intervene. I'd start with Cox for interpretability and if the proportional-hazards assumption fails, switch to random survival forests.

How would you validate these models before rolling them out?

Validation has to be temporal. I'd track C-index and time-dependent AUC, check the proportional-hazards assumption with Schoenfeld residuals, assess calibration against observed Kaplan-Meier curves, and evaluate fairness across customer segments.

And how would you test the retention interventions to determine if they were successful?

Each intervention should be A/B tested against a holdout. A contextual bandit can route different intervention types (discounts, service upgrades, outreach) to different segments.

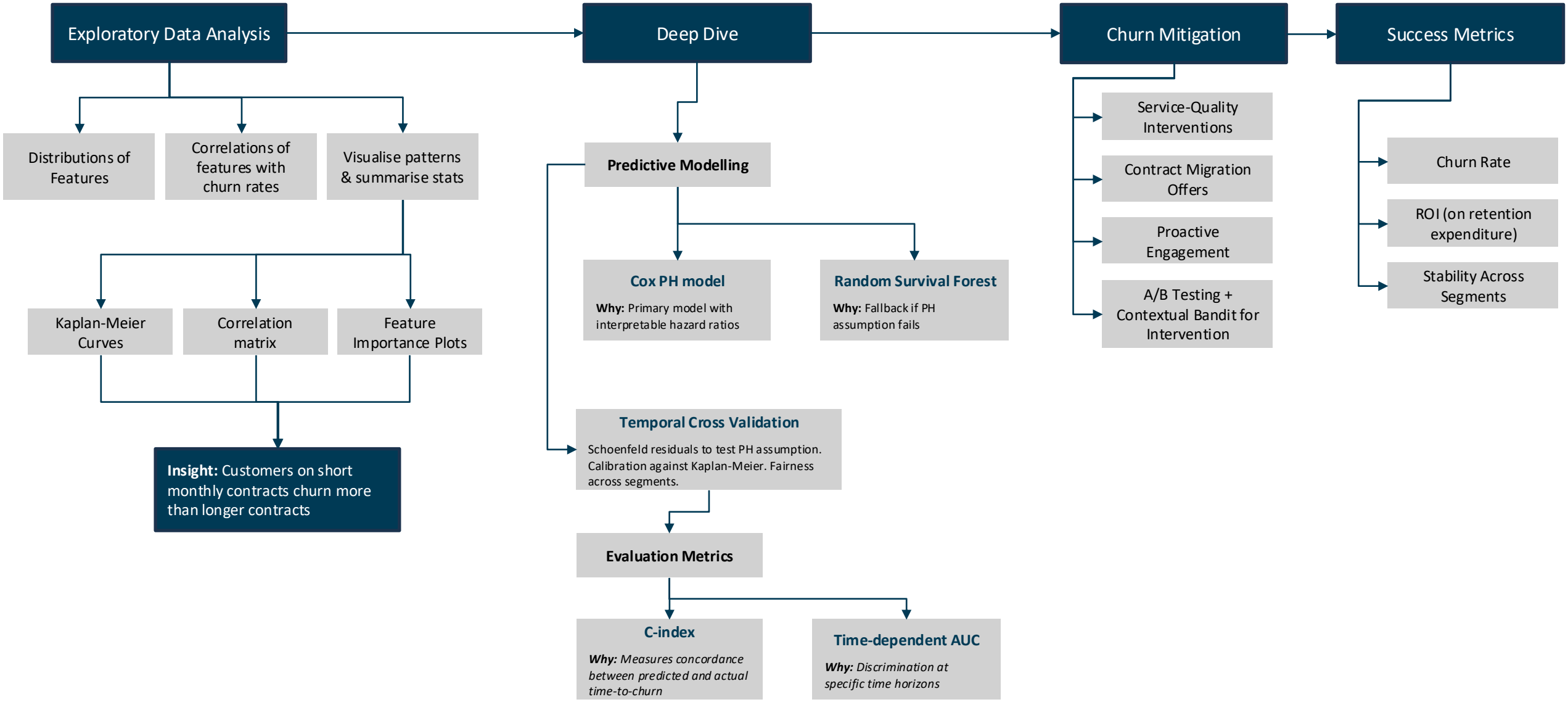
How would you measure if the programme is succeeding?

I'd track three things:

1. **Churn rate reduction** for treated cohorts versus control, particularly within the month-to-month segment, where the problem is concentrated
2. **ROI of retention spend**, since discounts and service upgrades have a real cost
3. **Stability across segments**, ensuring gains aren't concentrated in one geography or tenure band

Thank you for your comprehensive answers.

Reducing Telecom Customer Churn – Case Flow



Pricing a Pencil – Transcript (1/2)

Interviewer: Let's dive right into the case. Suppose a company that has been in the stationery business for a while is now entering the pencil segment. They've developed a pencil whose length never changes. Our task is to help them develop a pricing strategy.

Candidate: From what I understand, the client is already established in the stationery market, but this is their first venture into the pencil segment. Their USP is that the pencil's length never changes.

That's right. They are well-known in the stationery space but are new to making pencils. Now, considering this background, what would be your initial thoughts on developing a pricing strategy?

Given that it's their first time in the pencil market, I'd consider a few different approaches for the pricing strategy. Competitor benchmarking would be essential to understand the market rates. Additionally, I would look at cost-based analysis to ensure we cover our costs & maintain profitability. However, value-based pricing might be the most relevant here since the pencil's unique feature is that its length never changes, which could add significant value to the customer.

Can you explain how you would approach value-based pricing for this pencil?

Value-based pricing sets the price based on how much customers actually value the product. To estimate willingness-to-pay properly, I'd run a choice-based conjoint survey, i.e. show customers different pencil versions varying in durability, brand, and price, and ask them to pick. Their choices reveal how much each feature is worth in rupees. I'd analyse it using hierarchical Bayes (HB), which captures the full spread of WTP across customers as some will value durability heavily, while others won't. A single average would hide that.

I'd cross-check with a van Westendorp survey, which simply asks customers at what prices the pencil feels too cheap, too expensive, a bargain, or unaffordable. The output is a WTP range rather than a single point estimate.

Before we move to market sizing, can you give a rough numerical sense of where the WTP would land?

I'd look at three things.

1. **The substitute anchor –** standard pencils sell at ₹5 - 8, and premium variants at ₹15 - 20, so the premium-pencil category anchors at ~₹15
2. **The economic value –** If the forever pencil replaces ~50 standard pencils over five school years. years, the lifetime substitution value is ~₹250 (at ₹5 per pencil). But customers don't run lifecycle math. They anchor to familiar pencil prices and credit only a small fraction of the long-term savings, typically ~10% for everyday low-cost items, giving a realistic WTP of ~₹25
3. **Survey validation –** the conjoint HB posterior should peak around ₹25 with an interval of roughly ₹20–32, and van Westendorp's indifference point should sit just above the premium anchor

So, Durability premium = WTP (validated by surveys) – substitute anchor = ₹25 - ₹15 = ₹10

Good approach. The client is launching this product in Mumbai as a pilot before expanding pan-India. Target users are school-going kids, K through 5th grade. How would you estimate market size?

To estimate the market size, we need to know the number of school-going children in the target age group in Mumbai. One approach could be to use demographic data such as birth rates & population statistics. For example, if we know the number of births per year & the percentage of children attending school, we can estimate the total number of potential users.

That sounds reasonable. Let's say we have data that about 3 lakh births occur annually in Mumbai. How would you proceed with this information?

With 3 lakh births per year, we can estimate the number of children in the target age range by multiplying the birth rate by the number of years we are considering. For instance, if we are looking at children from ages 5 to 10, we would consider six age cohorts. So the estimated number of school-going children in that age group would be 3 lakh * 6, which equals 18 lakh children. If we assume that about 95% of children in Mumbai attend primary school, we have 18 lakh * 0.95 = ~17 lakh.

Good calculation. Now let's think about the usage. How many pencils do you think a school-going child uses annually, & what is the amount spent on pencils in those years?

From my understanding, school-going children use pencils extensively, especially in the early grades. Let's assume an average child uses about 10 pencils a year. At an average price of ₹5 per pencil, that's ₹50 per child per year. Multiplied across 17 lakh children, the addressable market in Mumbai is roughly ₹8.5 crores per year.

Because the forever pencil is consumed far less frequently, the annual market size may fall significantly, which means margins become more important than volume.

That's an important observation. Now we need to evaluate the cost structure. Suppose the variable cost of manufacturing is ₹15. How would you set the final price?

Given the ₹25 WTP we derived earlier and a variable cost of ₹15, the contribution margin is ₹10 per pencil. Before committing to a band, I'd build a demand model using XGBoost or LightGBM on store × week × price-tier features to project units sold under each candidate price, evaluated on WAPE. The

Pricing a Pencil – Transcript (2/2)

pilot then closes the loop by giving us actual elasticity to validate the model.

For the pilot, I'd test a tight band of ₹22 to ₹28 anchored around ₹25, leaning to the upper end, given the durability differentiator. This is a pilot test band, with room for upward revision. The long-term economic value is well above ₹25, and once the durability claim is validated, pricing can be revisited upward in phase two.

Excellent. Now let's delve into testing different pricing strategies in the pilot phase. How would you approach this?

Rather than charging different customers different prices in the same store, which raises fairness concerns, I'd do a geo-split, i.e. assign different prices to different groups of stores. Three prices (₹22, ₹25, ₹28) go to separate store clusters in Mumbai. To reliably detect sales differences between prices, we'd need around 40 - 60 stores per price point running for 8 weeks, long enough to cover normal buying patterns. I'd track customer complaints and retailer feedback as guardrails.

How would you analyse the test once it lands?

I'd use difference-in-differences (DiD) with pre-period sales as the control series to absorb store-level baseline variation. From the three price-volume points, I'd fit a log-log elasticity regression, i.e. $\log(\text{units}) = \alpha + \beta \cdot \log(\text{price})$, where β is the own-price elasticity. A hierarchical Bayesian version lets β vary by store cluster, revealing which segments the durability premium works in. Elasticity > 1 in magnitude means the price premium isn't justified by the durability story.

That's a great approach. Once the pilot launches, how would you focus on customer satisfaction and feedback?

For pricing-relevant signals specifically, I'd focus on three techniques:

1. **Sentiment Analysis –** A fine-tuned DistilBERT on labelled review data to track sentiment in reviews and social media, specifically around price and durability. A drop in sentiment when the price is discussed is an early signal of price resistance
2. **Topic Modelling –** BERTopic to surface common topics. For a new product, typical ones include perceived value, durability scepticism, and competitor comparisons
3. **Aspect-based feedback classification –** A fine-tuned ABSA model to classify each review by aspect (price, durability, usability), so the concerned team can act on each signal separately

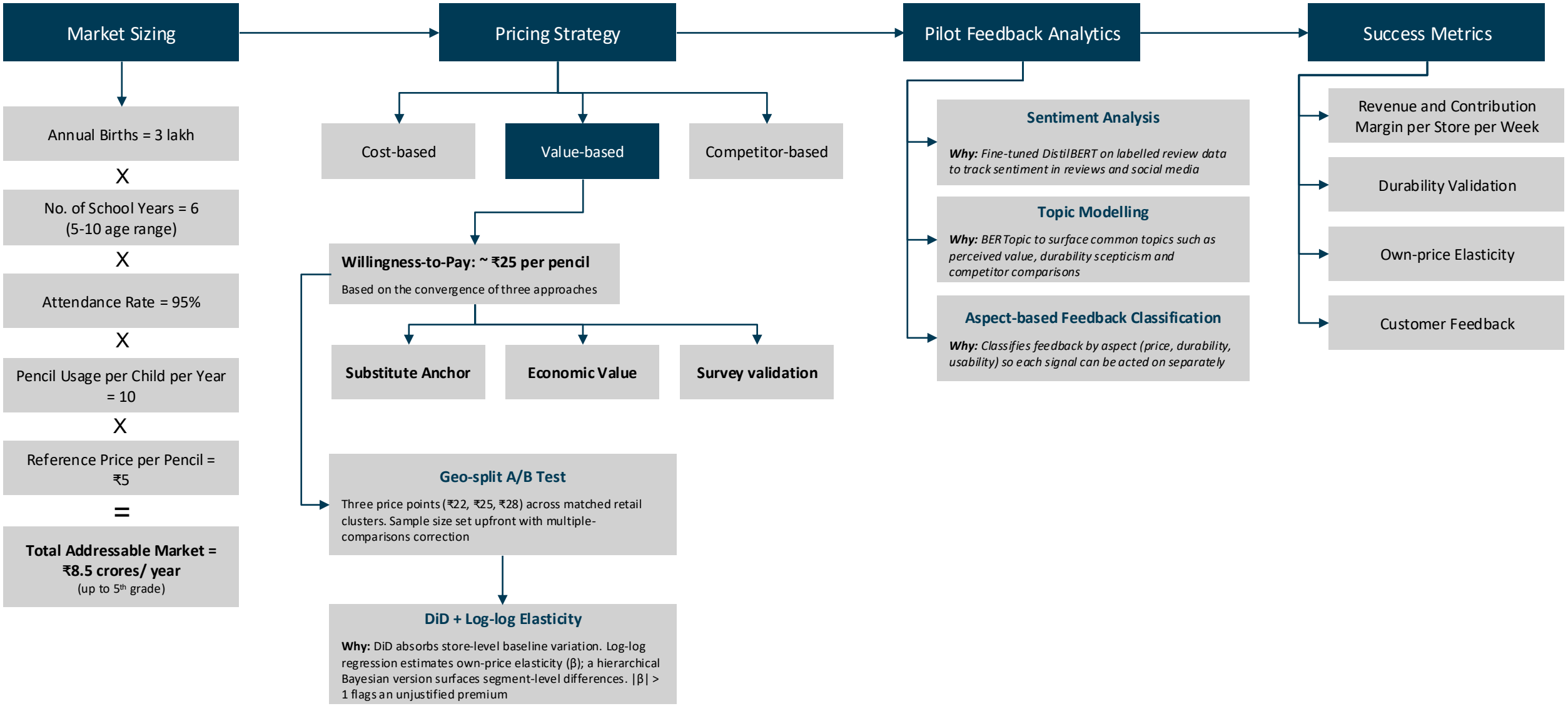
And how would you measure whether the pilot has succeeded enough to scale pan-India?

1. **Revenue and contribution margin per store per week** at the chosen price point, compared against a target derived from the addressable market estimate
2. **Durability validation,** i.e. % of customers still using the same pencil 6+ months post-purchase, since the value claim depends on long usage, not repeat buys
3. **Own-price elasticity** validates the durability premium. If |elasticity| > 1, the durability premium isn't earning its keep, and positioning needs revisiting before scale
4. **Customer feedback** signals strength on the durability claim, since this is what justifies any price premium over standard pencils

If all four clear thresholds set before the pilot, the case for pan-India scale-up is supported.

Those are valuable techniques. Thank you for your detailed analysis and insights. It was a pleasure discussing this case with you.

Pricing a Pencil – Case Flow



Automobile Industry

Industry Overview

The automobile industry covers the design, manufacturing, and sale of motor vehicles across two-wheelers, passenger cars, commercial vehicles, and EVs, along with components and aftermarket services.

India is the **3rd largest automobile market** and the **world's largest 2-wheeler market**, valued at **~$152B (2025)** and projected to reach **~$287B by 2033.** Maruti Suzuki, Tata Motors, Hyundai, Mahindra, Hero MotoCorp, Bajaj, Ashok Leyland, and TVS are the leading players.

MARUTI SUZUKI · TATA · Mahindra · Hero · BAJAJ · ASHOK LEYLAND · TVS

Business Segments

- **Vehicles**
  - 2 & 3 wheelers
  - Commercial vehicles
  - Passenger vehicles
  - Electric vehicles
- **Components**
  - OEM supply
  - Aftermarket parts
  - Tyres & Batteries
- **Services**
  - Dealership & Retail
  - Fleet Management
  - Auto Finance

Key Performance Indicators (KPIs)

Metric	Formula
Inventory Turnover Ratio	$\frac{\text{net revenue}}{\text{avg inventory value}}$
Production Efficiency	$\frac{actual\ production\ output}{\text{planned production } output} * 100$
Warranty Claim Rate	$\frac{no.\ of\ \text{warranty claims}}{total\ units\ produced} * 1000$
EV Penetration Rate	$\frac{\text{no. of EVs sold}}{\text{total vehicles sold}} * 100$
Dealer Throughput Rate	$\frac{\text{no. of vehicles sold per dealer}}{\text{total dealer capacity}} * 100$

Trends

1. EV transition accelerating with government targeting 30% EV penetration by 2030 through FAME III and PM E-Drive schemes
2. PLI scheme for auto components driving domestic manufacturing of advanced components
3. The SUV segment now accounts for over 50% of passenger vehicle sales
4. Software-defined vehicles gaining traction, with OEMs investing in telematics and OTA updates

Challenges

1. Heavy import dependence on China for EV components such as lithium cells, motors, and power electronics
2. Inadequate charging network and grid capacity slowing EV adoption outside metros
3. High regulatory transition costs due to simultaneous compliance with BS-VI Phase 2, CAFE norms, and upcoming safety mandates
4. Two-wheeler sales, which form the industry's volume base, are highly sensitive to monsoon and agricultural income cycles

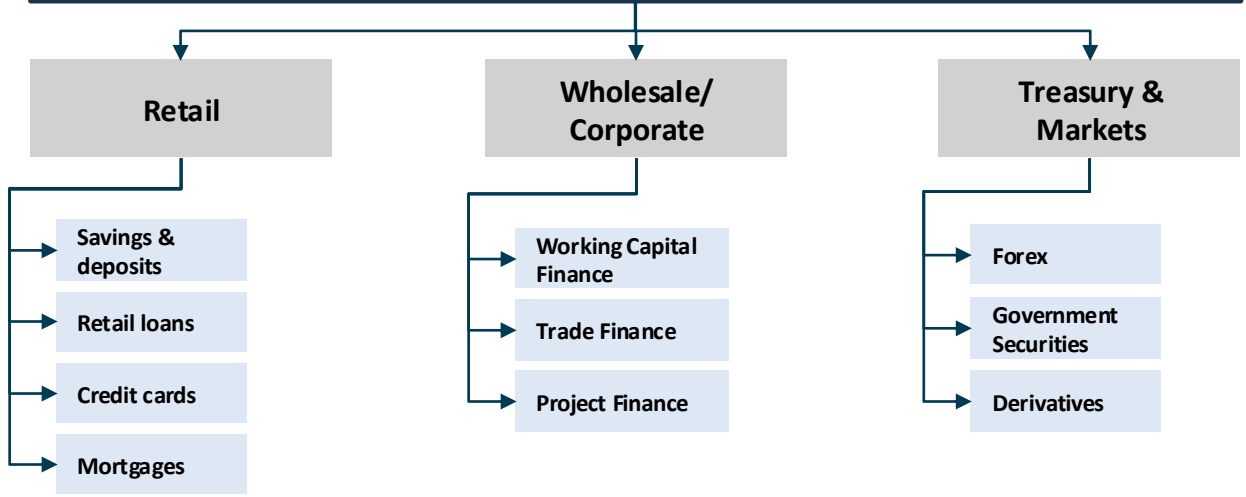
Banking Industry

Industry Overview

The banking industry encompasses the acceptance of deposits, extension of credit, and provision of financial services, including payments, wealth management, and insurance, serving retail, corporate, and government clients.

India's banking sector is valued at **~$3T in total assets** (2026)**.** The RBI regulates 12 public, 21 private, and 44 foreign banks. SBI, HDFC Bank, ICICI Bank, Axis Bank, Kotak Mahindra Bank, PNB, BoB, and IndusInd Bank are the leading players.

Business Segments



Key Performance Indicators (KPIs)

Metric	Formula
Net Interest Margin	$\frac{\text{interest income} - \text{interest expense}}{\text{avg interest} - \text{earning assets}} * 100$
Return on Assets	$\frac{\text{net profit}}{\text{avg total assets}} * 100$
Gross NPA Ratio	$\frac{\text{gross non} - \text{performing assets}}{\text{total advances}} * 100$
Credit-Deposit Ratio	$\frac{\text{total loans \& advances}}{\text{total deposits}} * 100$
Cost to Income Ratio	$\frac{\text{operating expenses}}{\text{net operating income}} * 100$

Trends

1. UPI and digital payments ecosystem processing over 19B transactions/month, reshaping retail banking revenue models
2. Co-lending partnerships between banks and NBFCs expanding credit access to underserved and rural segments
3. Neo-banking and embedded finance growing rapidly, with fintechs deepening into MSME and micro-credit segments
4. AI-driven credit scoring replacing traditional bureau-based models, particularly for first-time borrowers

Challenges

1. Elevated Gross NPA levels in public sector banks constraining fresh lending capacity despite recent improvements
2. Cybersecurity and fraud risk escalating with digital adoption, particularly in UPI and mobile banking channels
3. Priority sector lending compliance creating margin pressure, particularly for smaller private banks
4. Deposit mobilisation lagging credit growth, tightening liquidity and driving up cost of funds

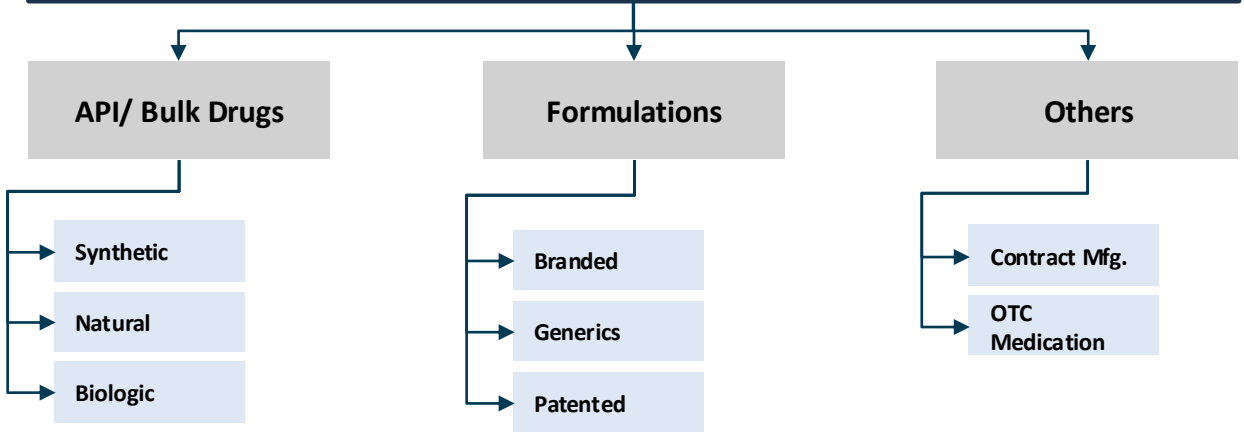
Pharmaceuticals Industry

Industry Overview

The pharmaceutical industry encompasses the R&D, manufacturing, and commercialisation of drugs, biologics, and vaccines.

India is the **3rd largest pharma producer by volume** globally, with over **~$60B** in annual sales, projected to reach **~$130B by 2030**. Abbott, Cipla, Zydus Wellness, Dr Reddy's, Sun Pharma, Sanofi, Lupin, and ManKind are the leading players.

Business Segments



Key Performance Indicators (KPIs)

Metric	Formula
Prescription Growth Rate	$\left(\frac{current\ period\ prescriptions}{previous\ period\ prescriptions} - 1\right) * 100$
Market Penetration Rate	$\frac{\text{no. of prescribers writing the drug}}{total\ no.\ \text{of prescribers}} * 100$
Patient Adherence Rate	$\frac{no.\ of\ patients\ adhering\ to\ treatment}{total\ no.\ of\ patients\ prescribed} * 100$
Clinical Trial Success Rate	$\frac{no.\ of\ successful\ trials}{total\ no.\ of\ trials\ conducted} * 100$
Adverse Event Reporting Rate	$\frac{no.\ of\ adverse\ events\ reported}{total\ units\ sold} * 1000$

Trends

1. Strategic shift from generics to biosimilars, biologics, and complex molecules
2. China+1 strategy driving global contract manufacturing toward India
3. Chronic disease surges, such as diabetes, oncology, and cardiology leading market growth
4. AI and digital platforms (e-pharmacy, GCCs) reshaping drug discovery and distribution

Challenges

1. ~70–72% API import dependence on China, structurally unresolved despite PLI schemes
2. NPPA price controls compressing margins on essential and chronic therapy segments
3. R&D spend (~7% of revenue) is far below global innovator benchmarks of 15–20%
4. Data integrity and GMP compliance failures causing costly export disruptions
5. Highly fragmented manufacturing base (~1,500 API facilities), limiting supply chain resilience

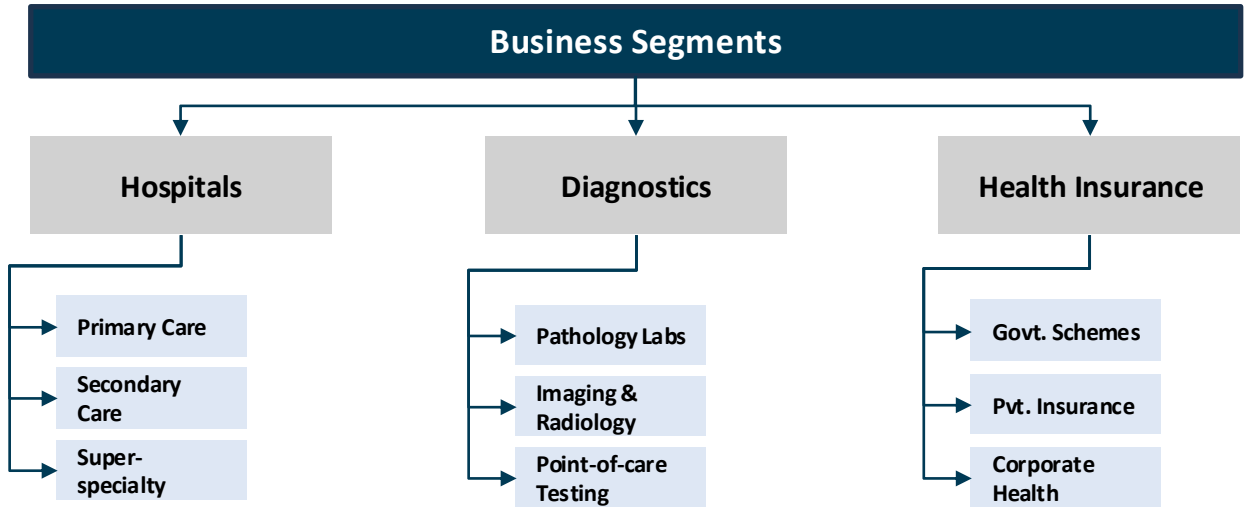
Healthcare Industry

Industry Overview

The healthcare industry encompasses medical equipment manufacturing and the delivery of preventive and curative care through hospitals, diagnostics, and insurance.

India's healthcare sector is valued at ~**$638B (2025)**. Apollo Hospitals, Fortis Healthcare, Max Healthcare, Narayana Health, Dr. Lal PathLabs, Thyrocare, Star Health Insurance, and Practo are the leading players

Business Segments



Key Performance Indicators (KPIs)

Metric	Formula
Avg Revenue per Occupied Bed	$\frac{\text{total inpatient revenue}}{\text{total occupied bed days}}$
Bed Occupancy Rate	$\frac{\text{occupied beds}}{\text{total available beds}} * 100$
Avg Length of Stay	$\frac{\text{total inpatient days}}{\text{total discharges}}$
Claims Settlement Ratio	$\frac{no.of\ \text{claims settled}}{\text{total claims filed}} * 100$
Patient Readmission Rate	$\frac{no.of\ \text{patients readmitted in 30 days}}{total\ \text{discharges}} * 100$

Trends

1. Ayushman Bharat PM-JAY is driving hospital network expansion into Tier 2 and 3 cities
2. Digital health mission (ABDM) building a unified health ID and interoperable medical records infrastructure
3. Indian medical tourism, valued at ~$12B and growing at ~12% CAGR
4. Health-tech investment surging, with telemedicine and AI-assisted diagnostics scaling rapidly post-pandemic
5. Chronic disease burden shifting hospital revenue mix toward long-term oncology, cardiology, and diabetes

Challenges

1. Out-of-pocket expenditure remains ~47% of total health spending, limiting access and driving catastrophic health costs
2. Health insurance penetration less than 40%, leaving a large underinsured population outside formal care networks
3. Regulatory fragmentation across CDSCO, NMC, IMA, and state health departments creating compliance complexity
4. Rural healthcare infrastructure severely underdeveloped

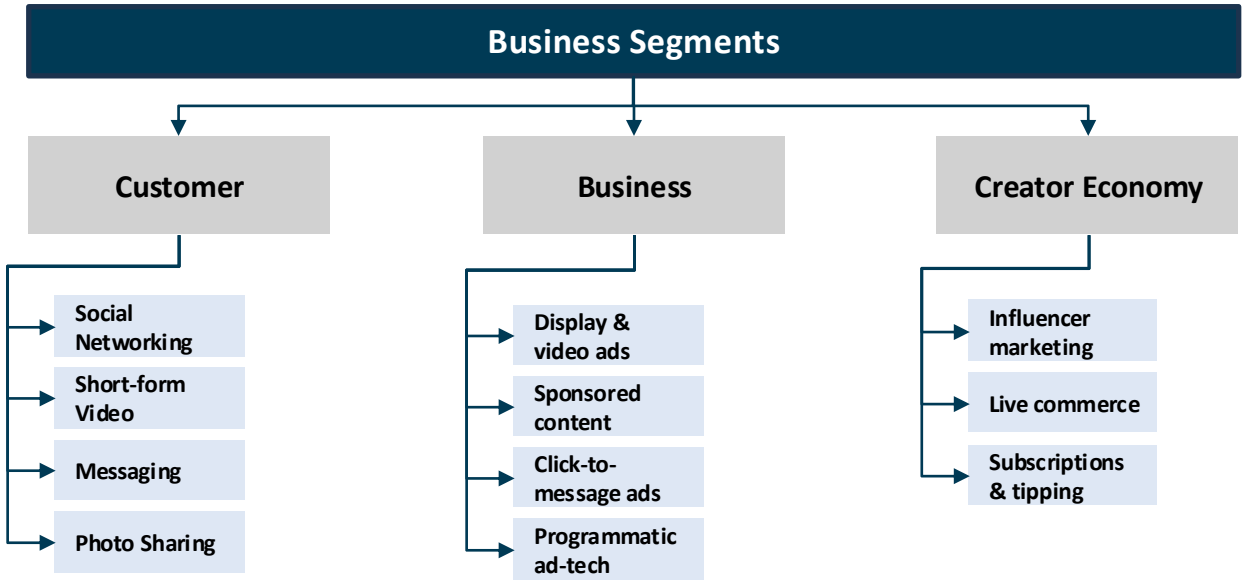
Social Media Industry

Industry Overview

Encompasses platforms enabling user-generated content, social networking, messaging, and short-form video, monetised primarily through advertising and premium subscriptions.

India is the **2nd largest social media market globally,** with **~491M active users (2025)** and **~$1.76B (2025)** in advertising revenue, projected to grow at **~15% CAGR**. Meta, Google, ByteDance, ShareChat, X, LinkedIn, Snapchat, and Telegram are the leading players.

Business Segments



Key Performance Indicators (KPIs)

Metric	Formula
DAU / MAU Stickiness Ratio	$\frac{\text{daily active users}}{\text{monthly active users}} * 100$
Average Session Duration	$\frac{\text{total session time}}{\text{total sessions}}$
Engagement Rate	$\frac{\text{likes} + \text{comments} + \text{shares}}{\text{total followers}} * 100$
Average Revenue Per User (ARPU)	$\frac{\text{total revenue}}{\text{avg users}}$
Cost Per Mille (CPM)	$\frac{\text{ad spends}}{\text{impressions}} * 1000$

Trends

1. Short-form video dominance, with Reels and Shorts, is reshaping content consumption and is now driving ~30% of digital ad spend in India
2. Tier-2/3 expansion, with regional-language platforms like ShareChat and Moj capturing audiences underserved by English-first global incumbents
3. India's influencer marketing spend is growing 25% YoY to ~$420M
4. AI-driven content recommendation and moderation, with platforms deploying ranking models, generative AI tools, and automated detection for unsafe content
5. Live shopping, click-to-message ads, and embedded checkout are blurring lines between social and e-commerce

Challenges

1. Evolving compliance burdens under India's IT Rules, including traceability mandates, the shift from 36-hour to 3-hour content takedown deadlines, and new deepfake labelling and watermarking obligations
2. Misinformation and deepfake pressure, with 24,000+ URLs blocked under Section 69A in 2025
3. User engagement saturation in Tier-1 markets, intensifying competition for time-on-app via algorithmic feeds
4. Rising SEBI and regulatory scrutiny on undisclosed influencer endorsements and AI-generated promotional content

Mobile Gaming Industry

Industry Overview

The mobile gaming industry covers the development, publishing, and monetisation of games designed for smartphones and tablets across casual, hardcore, real-money, and esports formats.

India is the world's **largest mobile gaming market by downloads**, valued at **~$3.7B (2024)** and projected to reach **$9.1B by 2029** at a **~20% CAGR**. Krafton, Dream11, Nazara, MPL, Gametion, Garena, Games24x7, and WinZO are the leading players.

Business Segments

- **Casual**
  - Hyper-casual
  - Puzzle & Arcade
  - Board Games
  - FPS
- **Real Money**
  - Fantasy Sports
  - Rummy & Poker
  - Cash Leagues
- **Esports**
  - Tournaments
  - Streaming & Broadcasting

Key Performance Indicators (KPIs)

Metric	Formula
Day-N Retention Rate (D1, D7, D30)	$\frac{\text{users active on day n}}{\text{new users on day 0}} * 100$
Average Revenue Per Daily Active User (ARPDAU)	$\frac{\text{total revenue}}{\text{daily active users}}$
Conversion Rate (Paying Users)	$\frac{\text{no. of paying users}}{\text{total active users}} * 100$
Lifetime Value (LTV)	$\text{ARPDAU} * \text{avg user lifespan}$
Customer Acquisition Cost (CAC)	$\frac{\text{total marketing and sales spend}}{\text{new users acquired}}$

Trends

1. In-app purchases are overtaking real-money gaming, with IAP growing 41% YoY and projected to surpass RMG revenue by FY29 at a 44% CAGR
2. Tier-2/3 expansion, with two-thirds of players now in non-metro cities and regional-language games
3. Esports formalisation, with the segment crossing $200M in 2024 and Krafton, Garena, and Nazara investing heavily in tournaments and college leagues
4. Cloud gaming and 5G adoption, with companies launching subscription-based platforms enabling console-grade play on budget devices
5. AI-driven personalisation for dynamic difficulty adjustment and personalised offer engines

Challenges

1. Online Gaming Act imposes a blanket ban on real-money games
2. 28% GST on deposits compressing margins for RMG platforms and accelerating consolidation toward cash-rich incumbents
3. Low ARPU at ~$3 versus mature markets
4. High CAC and steep early-stage churn, with most players dropping off within the first 30 days
5. Hardware and connectivity fragmentation across budget Android devices, raising QA costs and limiting graphics-intensive titles

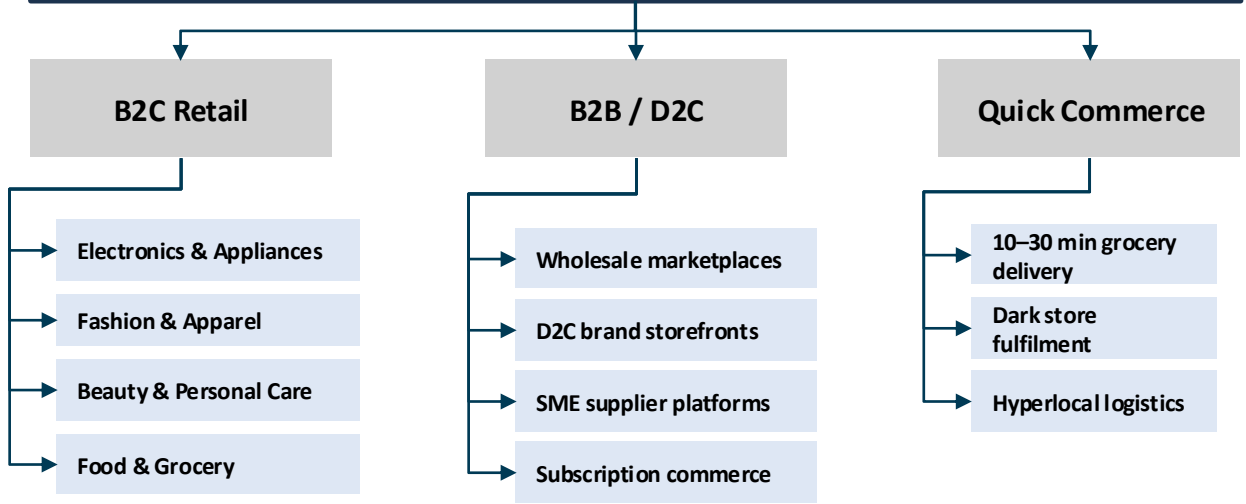
E-commerce Industry

Industry Overview

The e-commerce industry covers the buying and selling of goods and services over the internet, spanning B2C retail, B2B trade, quick commerce, and D2C brands.

India is the world's second-largest e-retail market by online shoppers, with over **270 million shoppers** (2024), and valued at **~$125B**, projected to reach **$345B by 2030**. Amazon, Flipkart, JioMart, Nykaa, Ajio, FirstCry, Meesho, and IndiaMart are the major players

Business Segments



Key Performance Indicators (KPIs)

Metric	Formula
GMV Growth Rate	$\left(\frac{\text{current period GMV}}{\text{previous period GMV}} - 1\right) \times 100$
Take Rate	$\frac{\text{revenue retained by platform}}{\text{total GMV}} \times 100$
Customer Acquisition Cost (CAC)	$\frac{\text{total marketing \& sales spend}}{\text{no. of new customers acquired}}$
Cart Abandonment Rate	$\left(1 - \frac{\text{no. of completed purchases}}{\text{no. of carts created}}\right) \times 100$
Return Rate	$\frac{\text{no. of orders returned}}{\text{total orders delivered}} \times 100$

Trends

1. Quick commerce expanding at 40%+ CAGR, with Zepto, Blinkit, and Swiggy Instamart now handling over two-thirds of all online grocery orders
2. Tier 2 and Tier 3 cities driving the next wave of growth, with rural online shoppers growing at ~22% CAGR
3. ONDC gaining traction as an open protocol, reducing seller dependence on large marketplaces and lowering commission costs
4. Social commerce and live shopping growing fast, with 80% of shoppers in smaller cities discovering products through social media

Challenges

1. Last-mile delivery in rural and semi-urban areas remains expensive and inconsistent due to poor road infrastructure and address ambiguity
2. High return rates, especially in fashion and apparel, inflate logistics costs and squeeze already thin margins
3. Cash on delivery still accounts for ~25% of orders in certain pin codes, complicating working capital cycles
4. Intense price wars are compressing profitability
5. Regulatory uncertainty around FDI norms and inventory-based vs. marketplace models continues to cloud long-term investment decisions

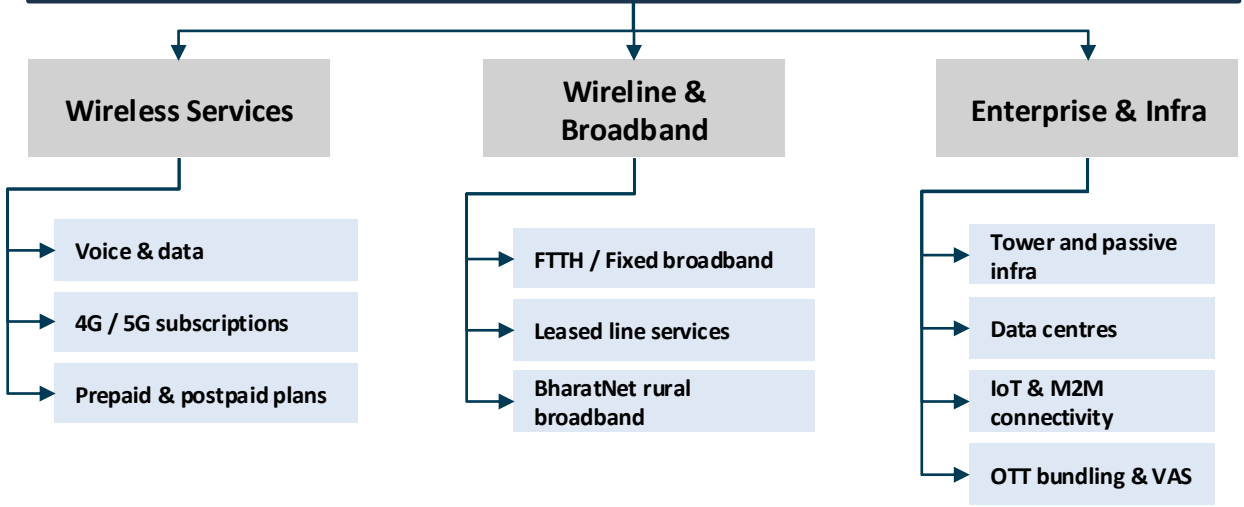
Telecom Industry

Industry Overview

The telecom industry covers the provision of voice, data, broadband, and enterprise connectivity services, along with infrastructure such as towers, fibre, and spectrum.

India is the world's **second-largest** telecom market with a subscriber base of **1.22 billion** and a tele-density of **86.09% (2025)**, valued at **~$43B.** Jio, Airtel, Vi, BSNL, Tata Communications, and Indus Towers are the leading players.

Business Segments



Key Performance Indicators (KPIs)

Metric	Formula
ARPU (Avg. Revenue Per User)	$\frac{\text{total revenue in a period}}{\text{average no. of subscribers in that period}}$
Data Usage Per User	$\frac{\text{total data consumed (GB)}}{\text{total active subscribers}}$
Churn Rate	$\frac{\text{no. of subscribers lost in a period}}{\text{total subscribers at start of period}} \times 100$
Call Drop Rate	$\frac{\text{no. of dropped calls}}{\text{total no. of calls initiated}} \times 100$
Revenue Market Share (RMS)	$\frac{\text{operator's revenue}}{\text{total industry revenue}} \times 100$

Trends

1. 5G rollout accelerating with India adding one of the fastest deployments globally; 5G subscriber base expected to hit 970 million by 2030 from 270 million in 2024
2. ARPU recovery underway after years of price wars, with Jio and Airtel both raising tariffs, improving industry profitability for the first time in a decade
3. Telecom companies are diversifying into adjacent areas such as fintech, cloud, OTT content, and enterprise SaaS to reduce dependence on pure connectivity revenue
4. Satellite broadband (LEO) is emerging as a new frontier, with Starlink and Jio SatCom targeting underserved rural and remote geographies

Challenges

1. Vodafone Idea's financial stress creates systemic risk. Its potential exit would further consolidate the market into a near-duopoly between Jio and Airtel
2. High spectrum auction costs and AGR dues continue to strain operator balance sheets
3. Only 46% of towers are currently fiberised, creating a bottleneck for full 5G deployment
4. Rural tele-density at ~60% lags urban areas significantly, and bridging this gap requires continued investment without proportionate revenue returns
5. OTT platforms riding on telecom pipes without contributing to network costs, a growing tension with regulators yet to find resolution

Important Statistical Tests (1/2)

Name	Statistic	Variable Description	Hypothesis	Explanation
Mean Comparison Tests				
Student's T–Test (equal variance)	$\frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	s_p: the standard deviation (equal) of sample 1 and 2	H_0: $\mu_1=\mu_2$ H_1: $\mu_1\neq\mu_2$	Compares the means of two groups assuming equal variances. Efficient under homoscedasticity but misleading if variances differ.
Welch's T–Test (unequal variance)	$\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$	s_1: the standard deviation of sample 1 s_2: the standard deviation of sample 2	H_0: $\mu_1=\mu_2$ H_1: $\mu_1\neq\mu_2$	Compares means without assuming equal variances. More robust in practice. Slightly less power if variances are actually equal.
One–Way ANOVA	$\frac{MS_{between}}{MS_{within}}$	$MS_{between} = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1}$ $MS_{within} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_1}(\bar{x}_{ij} - \bar{x}_i)^2}{N-k}$ k: the number of groups	H_0: $\mu_1=\mu_2=\ldots=\mu_k$ H_1: at least one $\mu_i\neq\mu_j$	Compares multiple group means through between-group vs within-group variance. Controls Type I error across many groups but does not identify which groups differ.
F–Test (overall model significance)	$\frac{SS_{reg}}{SS_{res}}\cdot\frac{n-k-1}{k}$	$SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$ $SS_{res} = \sum(y_i - \hat{y}_i)^2$ y_i, $\hat{y}_i$, $\bar{y}$: actual value, predicted value, mean of y_i k: number of features (excluding intercept)	H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$ H_1: at least one $\beta_j\neq0$	Compares explained vs residual variance to assess overall model significance. Indicates if any predictor is useful, but does not identify which one (use t-test for individual variable significance).
Proportion Tests				
Test of Proportions (one sample)	$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	$\hat{p}$: sample proportion p: claimed proportion	H_0: $p=\hat{p}$ H_1: $p\neq\hat{p}$	Compares an observed sample proportion to a hypothesized population proportion. Uses normal approximation under large samples to assess deviation from the benchmark. Inaccurate when sample size is small.
Test of Proportions (two samples)	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	$\hat{p}\,(pooled\ proportion) = \frac{x_1+x_2}{n_1+n_2}$ $\hat{p}_i$: sample proportion of the i–th sample	H_0: $p_1=p_2$ H_1: $p_1\neq p_2$	Compares proportions between two independent groups using a pooled estimate. Efficient for detecting differences in binary outcomes (e.g., A/B testing). Can be misleading when sample sizes are small or proportions are extreme.
Chi–Square (goodness of fit)	$\chi^2 = \sum\frac{(O-E)^2}{E}$	O: observed counts in sample E: expected counts if sample followed the specific distribution	H_0: variables follow expected dist. H_1: variables don't follow expected dist.	Compares observed vs expected categorical frequencies. Useful for distribution validation but requires sufficiently large expected counts.

x_i refers to the data points from the sample *$\bar{x}_i$ refers to the sample mean of $i-th$ sample* *$\bar{x}$ refers to the overall sample mean across all samples* *n_i: the number of data points in sample i* *n: the number of data points across all samples*

Important Statistical Tests (2/2)

Name	Statistic	Variable Description	Hypothesis	Explanation
Distribution Tests				
Shapiro–Wilk	$\frac{\left(\sum a_i \cdot x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$	$a = \frac{m^T \cdot V^{-1}}{\|V^{-1} \cdot m\|}$ $x_{(i)}$: i–th order statistic from the data m: expected values of sorted samples from normal dist. V: covariance matrix of the values in m	H_0: data is normally distributed H_1: data is not normally distributed	Correlates ordered sample values with expected normal quantiles; $W \approx 1$ indicates normality. Most powerful normality test for small to moderate samples (n < 5,000). Sensitive to slight deviations.
Kolmogorov–Smirnov (2–sample)	$sup_x \lvert F_n(x) - F(x) \rvert$	$F_n(x)$: empirical CDF from data F(x): theoretical CDF (e.g. normal)	H_0: data follows specified distribution H_1: data does not follow specified distribution	Measures the max gap between empirical and theoretical CDFs. Useful for general distribution comparison. Less powerful than the Shapiro-Wilk test for normality.
Jarque–Bera	$\frac{n}{6}\left(S^2 + \frac{(k-3)^2}{4}\right)$	S: sample skewness K: sample kurtosis	H_0: data is normally distributed H_1: data is not normally distributed	Tests normality using sample skewness and kurtosis. Detects deviations from symmetry and tail behaviour relative to a normal distribution. Simple and widely used in econometrics, but less powerful than the Shapiro–Wilk test for small samples.
Regression Diagnostics Tests				
Breusch–Pagan	$n \cdot R_\varepsilon^2$	R_ε^2: R sq from auxillary regression auxillary regression specification: $\hat{\varepsilon}_t^2 = \alpha + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_k X_{ki} + u_i$	H_0: $\gamma_1=\gamma_2=\ldots=\gamma_k=0$ H_1: at least one $\gamma_j \neq 0$	Regresses squared residuals on predictors to detect heteroskedasticity. Identifies non-constant variance. Assumes linear variance structure.
White's Heteroskedasticity	$n \cdot R^2$	R^2: R sq from auxillary regression auxillary regression specification: $\hat{\varepsilon}_t^2 = \alpha + \beta X_i + \gamma {X_i}^2 + \delta(X_i X_j) + u_i$	H_0: Homoskedasticity H_1: No Homoskedasticity (Heteroskedasticity)	Regresses squared residuals on predictors, their squares, and interaction terms to detect heteroskedasticity. Does not assume a specific variance structure, making it more flexible than Breusch–Pagan. Can suffer from overfitting and loss of power in small samples
Durbin–Watson	$\frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} {\hat{\varepsilon}_t}^2}$	$\hat{\varepsilon}_t$: residual at time t	H_0: $\rho=0$ (no first–order autocorrelation) H_0: $\rho \neq 0$	Measures correlation between consecutive residuals (DW ≈ 2 implies no autocorrelation). Detects serial dependence but is limited to first-order autocorrelation.
Breusch–Godfrey	$n \cdot R^2$	R^2: R sq from auxillary regression auxillary regression specification: $\hat{\varepsilon}_t = \alpha + \beta X_t + \rho_1 \hat{\varepsilon}_{t-1} + \rho_2 \hat{\varepsilon}_{t-2} + \cdots + \rho_p \hat{\varepsilon}_{t-p} + u_t$	H_0: $\rho_1=\rho_2=\ldots=\rho_p=0$ H_1: at least one $\rho_j \neq 0$	Regresses residuals on original predictors and their own lags to detect serial correlation. Identifies higher-order autocorrelation and works with lagged dependent variables. More flexible than Durbin–Watson, but relies on large-sample approximation and correct lag specification.

x_i refers to the data points from the sample *$\bar{x}_i$ refers to the sample mean of $i-th$ sample* *$\bar{x}$ refers to the overall sample mean across all samples* *n_i: the number of data points in sample i* *n: the number of data points across all samples*

Regression

A statistical method that models the relationship between a dependent variable (Y) and one or more independent variables (X), enabling prediction and causal inference.

TYPES

Linear

Fits a straight line, $y = X\beta + \in$ through the data. Finds the best line by minimizing SS_{res}.

Estimating housing prices, forecasting sales, etc.

Logistic

Applies a sigmoid function to the line and squeezes predictions between 0 and 1, outputting a probability. A threshold (usually 0.5) converts probability to class.

Predicting credit default risk, diagnosing diseases, etc.

Polynomial

Linear regression but with polynomial transformations to the feature space, X. Models non-linearity well, but is prone to overfitting.

Calculating returns to ad spends, drug dosage curve, etc.

Ridge/ Lasso

Handles multicollinearity through regularisation. The Ridge (L2) and Lasso (L1) penalty terms are $\lambda\Sigma\beta^2$ and $\lambda\Sigma|\beta|$.

Predicting house prices with numerous correlated features, identifying drivers of attrition from 100s of variables, etc.

Assumptions

Assumption	Explanation	Statistical Test
Linearity	The relationship between each X and Y is additive and linear *after* controlling for other variables. Violated when the true data-generating process is curved or multiplicative.	–
Normality of Errors	Residuals $\varepsilon \sim N(0, \sigma^2)$. Needed for valid p-values and confidence intervals in small samples. In large samples, CLT makes normality less critical.	Shapiro–Wilk test
Homoscedasticity	The spread of residuals is constant across all fitted values. Violation (heteroscedasticity) doesn't bias $\hat{\beta}$ but makes standard errors wrong, invalidating inference.	Breusch–Pagan test
No Autocorrelation	Residuals are independent of each other. Commonly violated in time series (today's error predicts tomorrow's). Causes underestimated standard errors.	Breusch–Godfrey test Durbin–Watson test
No Multicollinearity	Predictors are not near-linear combinations of each other. High VIF (> 10) inflates standard errors, making coefficients unstable and hard to interpret individually.	Variance Inflation Factor
No Endogeneity	X is uncorrelated with ε i.e., no omitted variables, reverse causality, or measurement error in X. This is the most serious violation; it biases $\hat{\beta}$ itself, not just the standard errors.	Hausman test

Important Concepts

Concept	Explanation
Dummy Variables	Categorical variables with k labels need $k-1$ dummies. The dropped category becomes the reference group and all coefficients are interpreted relative to it. Including all k creates perfect multicollinearity (dummy variable trap) because they sum to 1 identically.
Interaction Terms	Let $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2)$. Here, β_3 $(x_1 \cdot x_2)$ means the effect of X_1 on Y depends on the level of X_2. You cannot interpret β_1 and β_2 in isolation when an interaction is present. The marginal effect of X_1 is $\beta_1 + \beta_3 x_2$.
Residual Diagnostics	A fan shape means variance grows with $\hat{y}$ (heteroscedasticity → use robust SEs or log-transform Y); a curve means a linear model is mis-specified (add polynomial terms); clusters suggest a grouping variable was omitted.
Cross–Validation	In $k-fold\ CV$, data is split into $k\ folds$; the model trains on $k-1\ folds$ and tests on the held-out fold, rotating k times. Gives an unbiased estimate of out-of-sample error. Typical choices: $k = 5\ or\ 10$. Leave-one-out (LOO) is the extreme case $(k = n)$.
Degrees of Freedom	$df_{residual} = n - k - 1$ where $k\ =\ number\ of\ predictors$. Each estimated parameter "uses up" one df. Small df means less reliable variance estimates. This is why adding irrelevant predictors hurts. It costs df without reducing SS_{res} meaningfully.

OLS Regression Results – Model Fit Analysis


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.861
Model:                            OLS   Adj. R-squared:                  0.858
Method:                 Least Squares   F-statistic:                     270.9
Date:                Tue, 31 Mar 2026   Prob (F-statistic):          7.84e-127
Time:                        18:25:38   Log-Likelihood:                -783.48
No. Observations:                 313   AIC:                             1583.
Df Residuals:                     305   BIC:                             1613.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5459      5.188      0.105      0.916      -9.663      10.755
weight        -0.0203      0.002    -11.735      0.000***   -0.024      -0.017
weight_sq      2.3e-06   2.59e-07      8.867      0.000***  1.79e-06    2.81e-06
model_year     0.8311      0.053     15.636      0.000***    0.727       0.936
is_american   -1.6394      0.465     -3.529      0.000***   -2.554      -0.725
horsepower    -0.0196      0.013     -1.461      0.145      -0.046       0.007
cylinders      0.1539      0.245      0.627      0.531      -0.329       0.637
acceleration   0.0279      0.098      0.284      0.776      -0.165       0.221
==============================================================================
Omnibus:                       44.684   Durbin-Watson:                   1.957
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               87.134
Skew:                           0.769   Prob(JB):                     1.20e-19
Kurtosis:                       5.078   Cond. No.                     3.38e+08
==============================================================================
```

*** significant at 1% level ** significant at 5% level * significant at 10% level

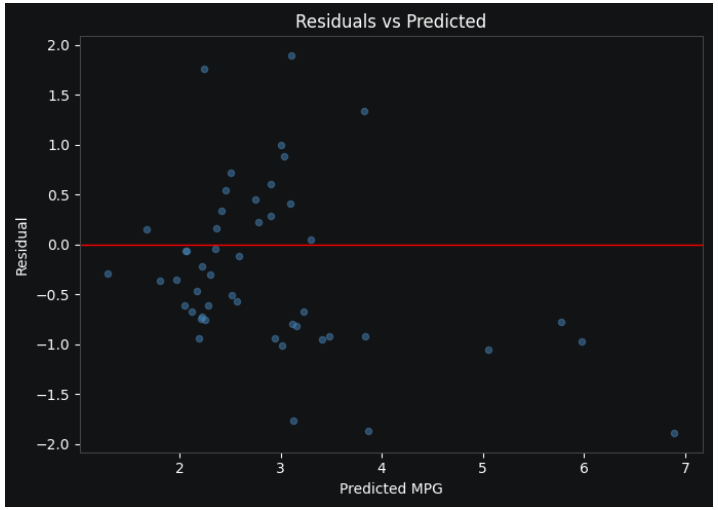
Metric	Formula	Explanation
β coef	$\Delta\hat{y}\ per\ unit\ \Delta X$	Estimated change in mileage for 1-unit increase in predictor X, holding others constant.
std err	$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}$	Uncertainty around the coefficient estimate. Smaller SE = tighter, more reliable estimate.
t−stat	$coef\ /\ SE$	Distance of β from zero in SE units.
P > |t|	$H_0: \beta = 0$ $H_1: \beta \neq 0$	Probability of observing $\|t\|$ this large under H_0. $p < 0.05 \rightarrow reject\ H_0$ (at 5% level of significance (α))
95% CI	$\beta_j \pm t_{1-\frac{\alpha}{2}} \times SE_j$	Confidence interval for the β coefficient. If the interval does not contain 0, the predictor is statistically significant at the 5% level. A narrow interval means a more precise estimate.
R−sq	$1 - \frac{SS_{res}}{SS_{tot}}$	86.1% of mileage variance explained by these features.
Adj. R−sq	$1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$	$R-sq$ but adjusted for number of predictors (7 in this case).
F−stat	$\frac{SS_{tot} - SS_{res}}{SS_{res}} \cdot \frac{n-k-1}{k}$	Joint significance of all predictors. Here: F = 270.9, p = 7.84e−127 → the 7 predictors jointly explain mileage far better than a model with no predictors.
Log−Likelihood	$-\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\hat{\sigma}^2)$ $-\frac{SS_{res}}{2\hat{\sigma}^2}$	Measures how well the model fits the data. A less negative value = better fit. Used in $AIC/\ BIC$ calculation.
AIC BIC	$2k - 2l$ $kln(n) - 2l$	Model comparison. Prefer the model with the lower AIC or BIC. BIC penalises model complexity more heavily, so it tends to favour simpler models.

Note: $SS_{tot} = \sum (y_i - \bar{y})^2$ $SS_{res} = \sum (y_i - \hat{y}_i)^2$ $SS_{tot} = SS_{res} + SS_{err}$

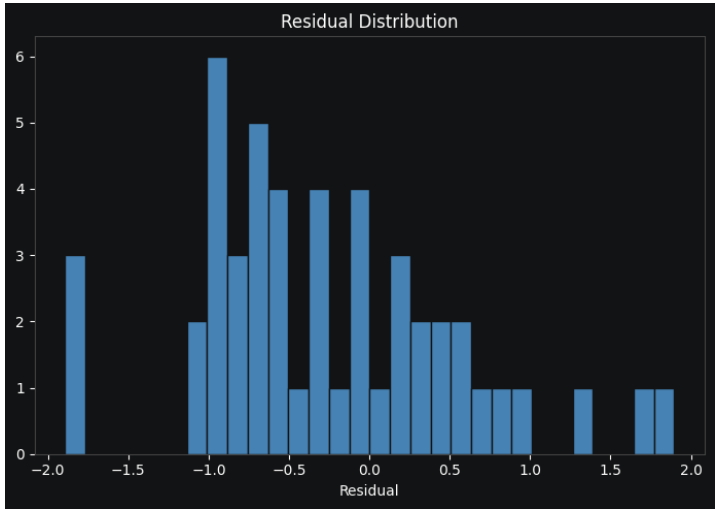
OLS Regression Results – Residuals Analysis


```
==============================================================================
Omnibus:                       44.684   Durbin-Watson:                   1.957
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               87.134
Skew:                           0.769   Prob(JB):                     1.20e-19
Kurtosis:                       5.078   Cond. No.                     3.38e+08
==============================================================================
```

Metric	Formula	Explanation
Omnibus	$\left(Z_1\left(\sqrt{b_1}\right)\right)^2 + \left(Z_2(b_2)\right)^2$	Combines normalized sample skewness and kurtosis into a single $\chi^2_{(2)}$ statistic; p < 0.05 rejects normality. Here, p = 0.000 implies residuals are not normal.
Skew	$\frac{1}{n}\left(\frac{e-\bar{e}}{\hat{\sigma}}\right)^3$	Measures asymmetry of the residual distribution. 0 = perfectly symmetric. Values between −1 and +1 are generally acceptable. Here: 0.769 → mild right skew.
Kurtosis	$\frac{1}{n}\left(\frac{e-\bar{e}}{\hat{\sigma}}\right)^4$	Measures tail heaviness. A normal distribution has kurtosis = 3. Higher values mean heavier tails (more outliers). Here: 5.078 → leptokurtic, heavy-tailed residuals.
Durbin–Watson	$\frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$	Tests for serial autocorrelation in residuals. Ranges from 0–4; values near 2 are safe. < 1.5 or > 2.5 is a concern. Here: 1.957 → acceptable.
Jarque–Bera	$\frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$	Normality test combining skewness and excess kurtosis. p < 0.05 → residuals not normal. Here: p = 1.20e−19 → clear non-normality.
Cond. No.	$\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$	Measures multicollinearity in predictors. > 30 = moderate concern; > 1000 = severe. Here: 3.38e+08 → extremely high, almost certainly due to weight and weight_sq being correlated.



Residuals are scattered randomly around zero with no clear pattern. The data points are limited, but it looks like the model has a tendency to overestimate mpg



The distribution is roughly centred around zero but shows a right skew and heavy tails, which is also confirmed by the *omnibus* metric

Classification

A supervised learning approach that assigns observations to predefined categories by learning a decision boundary from labelled training examples.

TYPES

Logistic Regression

Applies a sigmoid function to output class probabilities. The log-odds are modelled as a linear function of features. Extended to multiclass via softmax / OvR.

Used to predict whether a bank loan applicant will default.

Decision Tree

Recursively partitions the feature space by choosing splits that maximise information gain or minimise Gini impurity. Prone to overfitting; controlled via pruning and max-depth.

Used for diagnosing diseases from patient symptom profiles.

Random Forest

Ensemble of decorrelated trees trained on bootstrapped data samples with random feature subsets (bagging). Averages predictions to reduce variance.

Used for credit card fraud detection across millions of transactions.

Support Vector Machine (SVM)

Finds the maximum-margin hyperplane separating classes. Uses kernel functions (RBF, polynomial) for non-linear boundaries. Robust in high-dimensional spaces.

Real-life: Image recognition and spam e-mail filtering.

Confusion Matrix

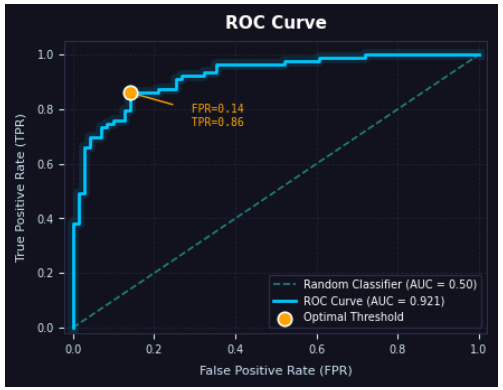
Unlike regression, which predicts continuous numeric values often summarised by a single error metric, classification requires a more granular evaluation because different errors carry different business costs.

A Confusion Matrix is the fundamental diagnostic tool for classification. It is an $n \times n$ matrix (where n is the number of class labels) that provides a cross-tabulation of actual versus predicted values. It serves as the building block for calculating key performance indicators like Precision, Recall, and the F1-Score.

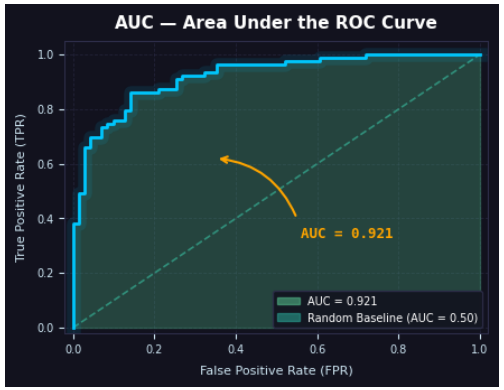
Predicted Labels \ Actual Labels	True	False
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

Metric	Formula	Explanation	Application
Accuracy	$\frac{TP+TN}{Total}$	The proportion of total predictions that were correct. Useful only when classes are balanced.	Standard quality checks where the cost of any error is relatively equal.
Precision	$\frac{TP}{TP+FP}$	Of all predicted positives, how many were actually correct? High precision minimises false alarms.	**Spam Detection:** An unflagged spam email is less costly than a flagged work email.
Recall/ Sensitivity (TPR)	$\frac{TP}{TP+FN}$	Of all actual positives, how many did we capture? High recall minimises missed cases.	**Medical Diagnosis:** Detecting a serious illness is more important than raising a false alarm on a healthy person.
Specificity (TNR)	$\frac{TN}{TN+FP}$	The proportion of actual negatives correctly identified.	**Drug Testing:** Ensuring a clean athlete does not test positive for a banned substance.
F1 Score	$\frac{2 \cdot prec \cdot recall}{prec+recall}$	The HM of Precision and Recall. It penalises extreme values of either, providing a balanced view of performance.	**Fraud Detection:** In banking, where you need to catch fraud without annoying customers through card blocks.

The **Receiver Operating Characteristic (ROC)** curve plots the **TPR** (y-axis) against the **FPR** (x-axis). It visualises the trade-off between catching positives and avoiding false alarms.



The **Area Under the Curve (AUC)** provides a single scalar value $(0\ to\ 1)$ to compare different models. It is the probability that the model will rank a randomly chosen positive instance higher than a negative one.



Classification Techniques

Logistic Regression

Models the probability of a binary outcome by mapping a linear combination of inputs to a value between 0 and 1 using the **Sigmoid function**.

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}$$

Strength: Highly interpretable as the coefficients directly represent the change in log-odds for a unit change in the predictor.

Weakness: Assumes a linear relationship between independent variables and the log-odds of the dependent variable, which may not always be true.

Used in predicting probability of loan default in credit risk modelling or calculating Alpha probability in stock movement.

Decision Tree

A non-parametric model that recursively splits the data into subsets based on feature values to maximise the homogeneity (purity) of the resulting nodes. The splitting can be done by minimising the $Gini\ Index$ or $Entropy$ at each node.

$$Gini\ Index = 1 - \sum_{i=1}^{c} p_i^2 \qquad Entropy = -\sum_{i=1}^{c} p_i \cdot log_2(p_i)$$

Strength: Easy to visualize and explain to stakeholders; requires minimal data preprocessing (no scaling needed).

Weakness: High variance; small changes in data can lead to a completely different tree structure (overfitting).

Used for identifying critical fail points in supply chain logistics or initial feature selection for complex datasets.

Random Forest

An ensemble method that builds multiple decision trees using **Bootstrap Aggregation (Bagging)** and selects the final class based on a majority vote.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

Strength: Exceptionally robust to noise and outliers; significantly reduces the risk of overfitting compared to a single tree.

Weakness: Increased computational cost and memory usage; loses the direct interpretability of a single decision tree.

Used for high-accuracy fraud detection in financial transactions or predicting disruptions in global supply networks.

Support Vector Machine (SVM)

Finds the hyperplane that maximises the margin between classes in a high-dimensional space.

$$min \frac{1}{2} ||w||^2 \ subject\ to\ y_1(w \cdot x_i + b) \geq 0$$

Strength: Effective in high-dimensional spaces and versatile through the use of different **Kernels** (Linear, RBF, Polynomial).

Weakness: Sensitive to the choice of kernel and regularization parameters; does not directly provide probability estimates.

Used in text-based sentiment analysis for corporate governance reports or image-based defect detection in manufacturing.

Clustering

An unsupervised learning technique that groups unlabelled observations so that points within a cluster are more similar to each other than to points in other clusters. Similarity is typically defined using a distance metric such as Euclidean, Manhattan, Cosine, etc.

Types

K-Means

Partitions data into exactly K clusters (can be decided using the elbow method).Iteratively assigns each point to its nearest centroid, then recomputes centroids until convergence. Sensitive to initialisation;

Used for customer segmentation for targeted marketing campaigns.

DBSCAN

No need to pre-specify K. Groups points in dense regions; labels sparse outliers as noise. A point is a *core point* if it has ≥ MinPts neighbours within radius ε. Clusters grow by adding density-reachable points. Can discover non-convex, arbitrary shapes.

Used for detecting anomalies and outliers in network traffic logs.

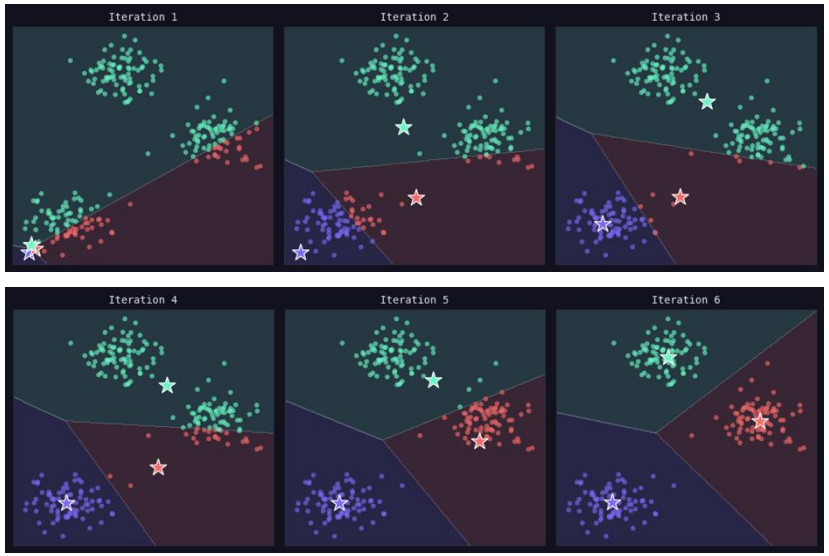
Hierarchical Clustering

Builds a tree of nested clusters (dendrogram) via agglomerative (bottom-up) merging. Cut the tree at a chosen height to yield K clusters. Key concepts: linkage criteria (single, complete, Ward), dendrogram, cophenetic correlation.

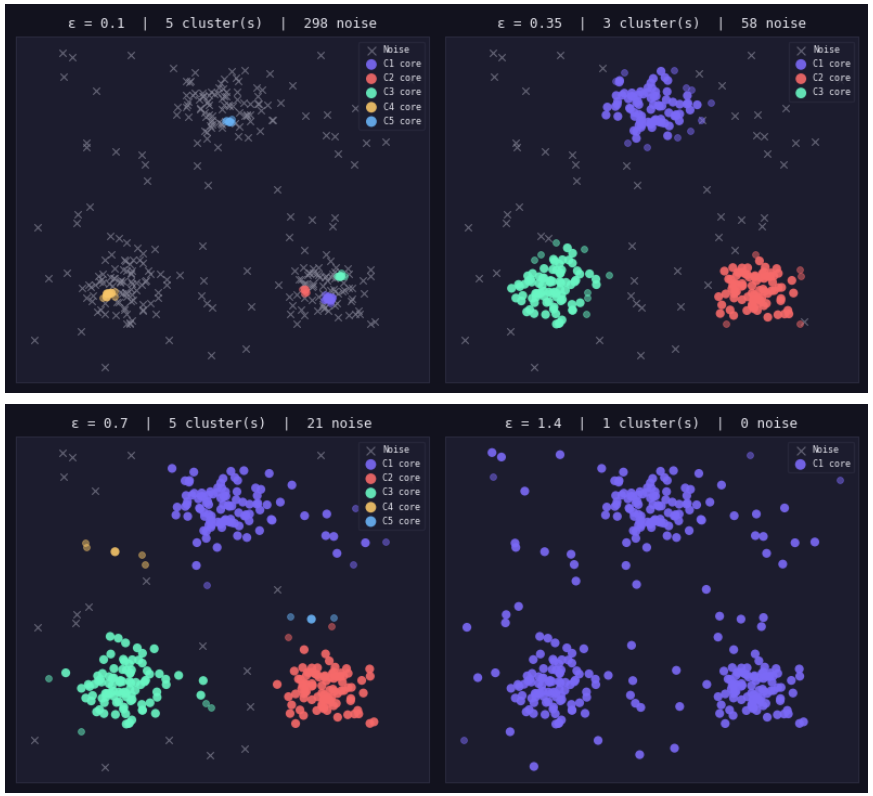
Used for grouping gene expression profiles in genomics research.

Key Metrics

Concept	Formula	Explanation
Inertia	$\sum_{i=1}^{k} \sum_{x \in C_i} \|\|x - \mu_i\|\|^2$	The sum of squared Euclidean distances from each point (x_i) to its assigned centroid (C_i), having the coordinates μ_i. K-Means directly minimises this. Lower inertia = tighter clusters. But inertia always decreases as K increases, so it cannot alone determine the optimal K. Always standardise features as inertia is scale-sensitive.
Silhouette Score	$\frac{b(i) - a(i)}{max(a(i), b(i))}$	Measures how well each point fits its own cluster vs. the nearest other cluster. a_i is the average distance between i and all the other points in its own cluster and b_i is the distance between i and the nearest next cluster centroid. Range: [-1, +1]. Score near +1 → well separated; near 0 → on the boundary; negative → possibly misclassified. Average across all points gives an overall cluster quality metric. Works for any algorithm (unlike inertia).



While 3 clusters are visually obvious in this simple example, real-world high-dimensional data is rarely this clear. In such cases, the elbow method can be used to identify the optimal number of clusters by finding the point where adding more clusters yields diminishing returns in inertia. K-Means is also notably sensitive to initialisation. Different starting centroids can lead to vastly different final boundaries.



DBSCAN is primarily used for high-dimensional data where clusters have arbitrary, non-spherical shapes. The algorithm is highly sensitive to its parameters (ε and $MinPts$). As the subplots demonstrate, a value that is too small treats the majority of the data as noise, while a value too large merges distinct groups into a single cluster.

Time Series

A sequence of observations indexed in time order. Modelling exploits autocorrelation and trend to forecast future values or understand temporal dynamics.

Types

AR (Auto Regressive)

Current value is a linear combination of its own past p values. Captures momentum and mean-reversion patterns.

Used for stock prices, economic indicators, etc.

MA (Moving Average)

Models current value as a linear combination of past q error (shock) terms. Good for smoothing and short-run noise.

Used for demand forecasting, inventory smoothing, etc.

ARIMA

Combines AR, differencing (I), and MA. The d parameter makes non-stationary series stationary. Most general classical TS model.

Used for GDP forecasting, inflation modelling, etc.

SARIMA/ Prophet

SARIMA adds seasonal differencing and seasonal AR/MA terms. Prophet decomposes into trend, seasonality and holiday effects.

Used for retail sales, web traffic with seasonality, etc.

Assumptions

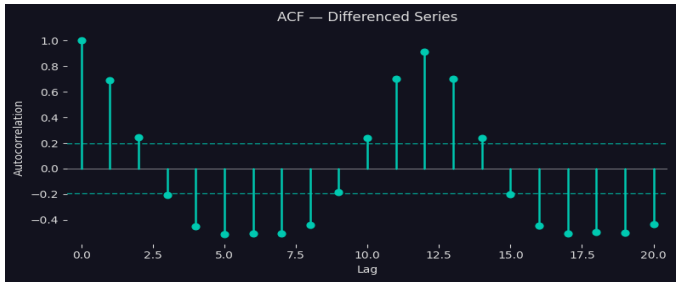
Assumption	Explanation	Statistical Test
Stationarity	Stationarity requires that the mean, variance, and autocovariance are constant over time. If a series is non-stationary, it must be differenced or log-transformed before modelling.	ADF / KPSS **test**
No Autocorrelation	Residuals after fitting should be white noise, i.e. zero mean, constant variance, and no autocorrelation at any lag. If the residuals are not white noise, then the model has not captured all the temporal structure.	**Ljung-Box test**
Linearity	The current value is a linear function of past values (AR) or past errors (MA).	-
Normality of Errors	Residuals follow a normal distribution with zero mean. Required for valid inference; less critical with large samples due to the CLT.	**Shapiro-Wilk test**
No Structural Breaks	The data-generating process is stable over the estimation window. Regime changes or policy shifts violate this assumption.	**Chow /** CUSUM **test**
Homoscedasticity	Error variance is constant over time. Volatility clustering in financial series violates this; use $ARCH/\ GARCH$ to model time-varying variance.	ARCH−LM **test**
Exogeneity of Regressors	In $ARIMAX$, external regressors must be uncorrelated with the error term. Violation biases coefficient estimates, as in the OLS endogeneity problem.	**Hausman test**

Autocorrelation Function (ACF)

Total correlation between y_t and y_{t-k}, including lagged effects

$$\rho(\mathrm{k}) = \frac{\gamma(k)}{\gamma(0)}$$

where, $\gamma(k) = \sum_{t=k+1}^{n}\left((y_t - \bar{y}) \cdot (y_{t-k} - \bar{y})\right)$ and $-1 \leq \rho(\mathrm{k}) \leq 1$

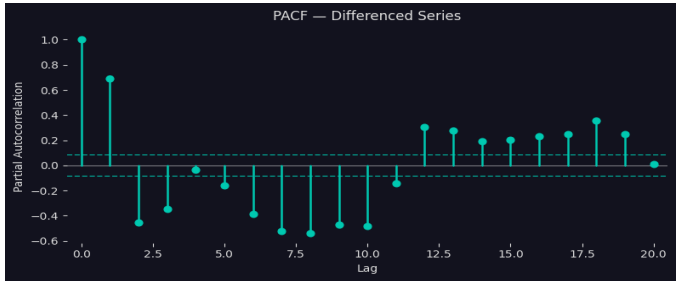


ACF plot shows some kind of seasonality in the differenced series. A $SARIMAX$ model might fit this data well.

Partial Autocorrelation Function (PACF)

Direct correlation between y_t and y_{t-k}, after removing intermediate lagged effects

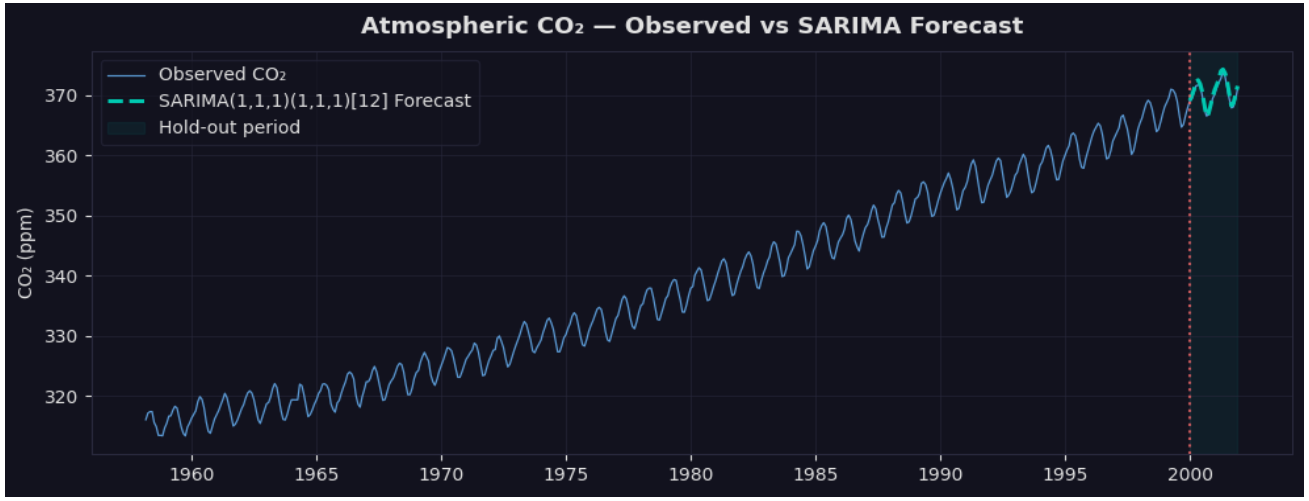
$$\varphi(\mathrm{k},\mathrm{k}) = \frac{[\rho(k) - \sum \varphi(\mathrm{k}-1, j) \cdot \rho(k-j)]}{[1 - \sum \varphi(k-1, j) \cdot \rho(j)]}$$



$PACF$ plot also confirms the findings in the ACF plot.

Time Series (SARIMA) – Model Fit Analysis


```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                            log_co2   No. Observations:                  502
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                2610.568
Date:                            Wed, 01 Apr 2026   AIC                           -5211.136
Time:                                    02:15:48   BIC                           -5190.319
Sample:                                03-01-1958   HQIC                          -5202.949
                                     - 12-01-1999
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1607      0.124      1.300      0.194      -0.082       0.403
ma.L1         -0.4823      0.110     -4.386      0.000      -0.698      -0.267
ar.S.L12      -0.0862      0.061     -1.422      0.155      -0.205       0.033
ma.S.L12      -0.6069      0.068     -8.981      0.000      -0.739      -0.474
sigma2      9.75e-07   5.51e-08     17.679      0.000    8.67e-07    1.08e-06
==========================================================================================
Ljung-Box (L1) (Q):                   0.37   Jarque-Bera (JB):                 42.54
Prob(Q):                              0.55   Prob(JB):                          0.00
Heteroskedasticity (H):               0.50   Skew:                             -0.14
Prob(H) (two-sided):                  0.00   Kurtosis:                          4.44
==========================================================================================
```



Test of Stationarity

		ADF Test	KPSS Test	Verdict
Before Differencing	Statistic	1.9254	3.3796	Fail to reject H_0 in ADF – non-stationary
	P–value	(p=0.9986)	(p=0.0100)	Reject H_0 in $KPSS$ – non-stationary
After Differencing	Statistic	−5.1177	0.2318	Reject H_0 in ADF – stationary
	P–value	(p=0.0000)	(p=0.1000)	Fail to reject H_0 in $KPSS$ – stationary

Model Selection

Choose best model based on lowest AIC / BIC. Here, after doing a $grid\ search$ over p, d, q values, we choose $SARIMA(1,1,1)(1,1,1)[12]$

Evaluation Metrics

Forecast Accuracy	
MAPE	0.14% ppm
RMSE	0.600 ppm

Residual Analysis		
Ljung–Box	p=0.55	Residuals are white noise (no autocorrelation)
Jarque–Bera	p=0.00	Residuals are not normally distributed

Advanced Techniques

Technique	Explanation
ARCH	Models volatility clustering by assuming that the variance of the current error term depends on the sizes of previous error terms.
GARCH	Extends ARCH by incorporating past forecast variances along with past squared errors, allowing for a more flexible way to model long-term persistence in volatility.
VAR	Captures the linear interdependencies among multiple time series by treating every variable as a function of the past values of itself and all other variables in the system.
Cointegration	A statistical property where two or more non-stationary time series share a common long-term equilibrium. Their linear combination is stationary even if the individual series drift apart in the short run.

Neural Networks

A computational framework of interconnected nodes organised in layers that learns complex, non-linear by iteratively adjusting weights through backpropagation and gradient descent.

Types

Feedforward Neural Network (FNN)

Data flows in one direction from input → hidden layers → output. Each neuron applies a weighted sum followed by a non-linear activation (ReLU, sigmoid, tanh).

Used for credit scoring, pricing models, etc.

Convolutional Neural Network (CNN)

Uses convolutional filters to detect spatial hierarchies. Pooling layers reduce dimensionality while retaining key features. Highly parameter-efficient for image data.

Used for image classification, medical imaging, facial recognition, etc.

Recurrent Neural Network (RNN) / LSTM

Processes sequential data by maintaining a hidden state passed across time steps. LSTMs add gating mechanisms (input, forget, output gates) to resolve the vanishing gradient problem and capture long-range dependencies.

Used for text generation, speech recognition, etc.

Transformer

Replaces recurrence with self-attention, computing pairwise relationships across all tokens simultaneously. Positional encodings preserve sequence order. Scales efficiently and underpins modern large language models.

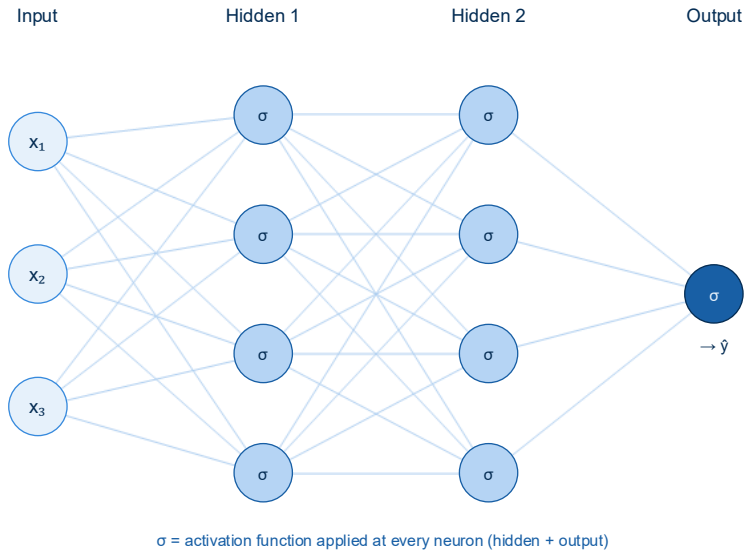
Used for large language models, code generation, etc.

Important Concepts

Concept	Explanation
Input Layer	The first layer that receives raw feature data. Each node corresponds to one input variable, and there is no computation at this layer. The number of nodes equals the number of features in the dataset.
Hidden Layers	Intermediate layers between input and output, where the actual learning happens. Each neuron computes a weighted sum of its inputs, adds a bias term, and passes the result through an activation function. Deeper networks with more hidden layers can capture increasingly abstract patterns.
Output Layer	Produces the final prediction. The number of nodes and the activation function depend on the task. A single node with sigmoid can be used for binary classification, multiple nodes with softmax can be used for multiclass, or a single linear node can be used for regression.
Activation Function	Introduce non-linearity, allowing the network to learn complex patterns. Without them, stacking layers collapses into a single linear transformation. Commonly used activation functions are ReLu, Sigmoid, Softmax, and Tanh.
Backpropagation	After a forward pass generates a prediction, the error is computed via a loss function, and the parameters are adjusted through backpropagation. Backpropagation applies the chain rule to calculate the gradient of the loss with respect to every weight in the network, moving backwards layer by layer. These gradients are then used by an optimiser to update weights.
Vanishing Gradient	As gradients are propagated backwards through many layers, repeated multiplication of small values causes them to shrink exponentially, effectively vanishing, which makes deep networks hard to train. Possible solutions include ReLU, batch normalisation, and residual connections.
Dropout	A regularisation technique where a random fraction of neurons are set to zero during each training step. This prevents neurons from co-adapting and forces the network to learn more robust, distributed representations. Typically applied only during training, not inference.

Activation Functions

Function	Formula	Explanation
ReLu	$\max(0, x)$	Outputs the input if positive, zero otherwise. Simple and computationally efficient. Can suffer from "dying ReLU" where neurons get stuck outputting zero permanently.
Sigmoid	$\frac{1}{1+e^{-x}}$	Squishes input to $\mathbb{R}[0, 1]$. Interpreted as a probability, making it a binary classifier. Prone to vanishing gradients at extreme values.
Softmax	$\frac{e^{x_i}}{\sum e^{x_j}}$	Converts a vector of raw scores into a probability distribution. Used exclusively in the output layer for multiclass classification.
Tanh	$\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	Similar shape to sigmoid but outputs $\mathbb{R}[-1, 1]$. Zero-centred, making gradient updates more stable than sigmoid but continues to suffer from saturation at extremes.



Scan this link and let us know what you think of the casebook
(or just click on the QR code).