*Management is thinking of launching an instalment payment financial product for people who would like to purchase a more expensive product and pay it in equal instalment payments over a period of 12/18/24 months.*

**Q1.1 How would you model the risk of each user when no historical credit data is available on that user for month 0?**

Acquire Credit Data

* Seek consent. Gojek can seek user consent to access credit history from third-parties like the Commercial Credit Bureau, in return for loyalty points/vouchers.
* Buy/Use Third-party data. Certain agencies offer additional data e.g, Commercial Credit Bureau (default events), ACRA (nationality, residence).

Use Non-instalment/Non-Credit Data

Acquire related data, and use them if correlations with default rates of long-term users are good.

* Existing Internal Data. Gojek has existing data like transactions (payment history), demographics (age), behaviour (customer rating)
* Government Data. In Singapore, ‘MyInfo’ offers demographics, personal finance, employment & education, vehicle, and property data. e.g. CPF contribution is 20% of monthly gross pay, a useful proxy for income

Use Bayes’ Theorem

* Calculate the probability of default, based on prior probabilities of events (calculated from data mentioned above) related to defaults. E.g. priors like customers between ages 20-30 repay 70% of the time. Use prior probabilities to calculate the posterior probabilities of default.

Unsupervised Learning

* Build a clustering model on existing users without instalment/credit data (same variables as new users), group by clusters, find the mean risk and default rate. Predict clusters for new users, and attach the corresponding risk and default rate.

[198 words]

**Q1.2 How would you model the risk of each user when first/second/third month repayment data is available?**

Update Bayesian’ Model

* This risk model can be updated, and made more accurate, with new instalment history to account for systematic differences between Gojek users

Granularized Time Periods

* A time period of 1-3 months may not give us enough instalment data to accurately model default risk. However, if the instalment plan allows for flexible ‘repay anytime’ schedules e.g. weekly basis, these may give us enough instances to create categories for ‘prompt payment’, ‘prompt-slow payments’, and ‘slow payments’. We can then engineer additional features e.g. % slow payment, average payment intervals.

Develop Separate Predictive Model for New Users

* Use existing users’ data with first/second/third month repayment variables to see if they are good predictors of subsequent months’ repayment, and for how many of those X months. If they are, then a separate predictive model for new users can be built to predict default risk, but used only to predict for X more months. After which point, we should revert to the standard risk model used for existing users.

[170 words]

*Management noticed that the output of your risk model (range from 0-1 for probability of default) is different from the actual default rates. For example, users with the risk score of 0.1 and below default at 20% rate and users with the risk score 0.9 and above default 60% of the time.*

**Q2.1 How would you align the output of your credit scoring model with the actual default rates?**

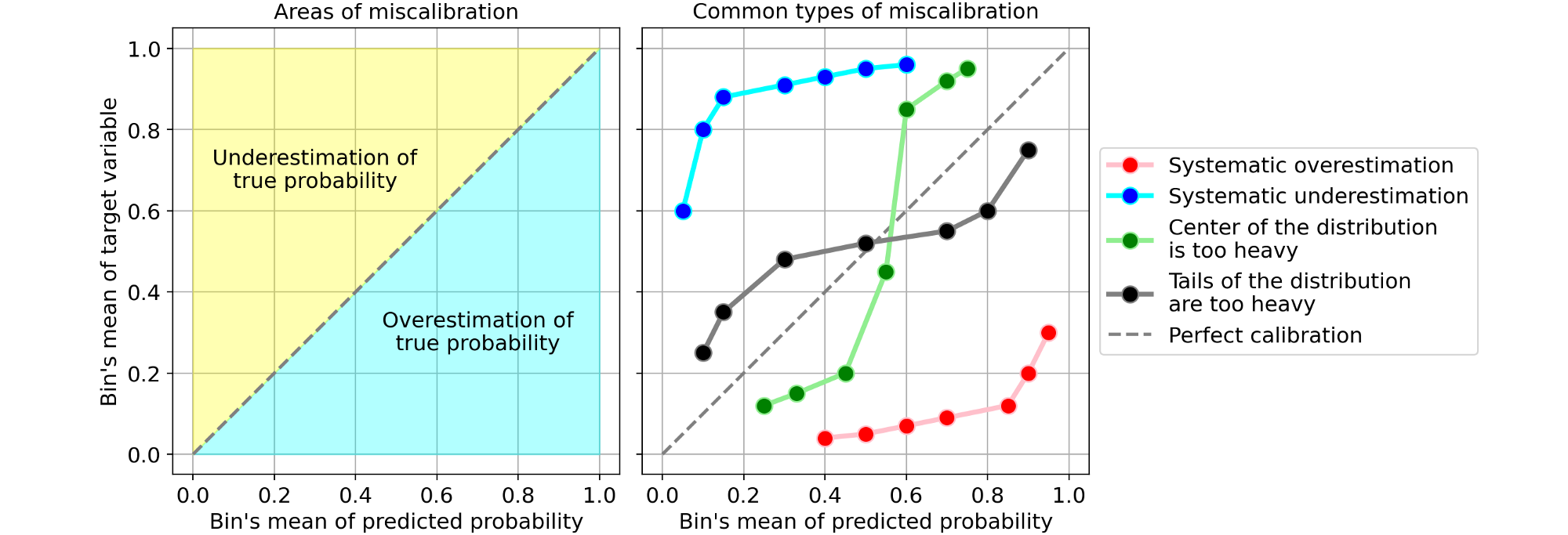
The problem is that credit scores are not directly proportional to the probability of failure, which affects explainability and usability.

Diagnosing Underlying Issues

* Determine if the model's predictive power is the problem. If it isn't performing well, this may account for the uneven ratio of credit score to default rate. The model will have to be revisited.
* Determine if high default probability users are being eliminated beforehand lowering the probability ceiling for the poor credit scores

Solution

* Regardless, model calibration can help to rescale the probability distribution. Classification algorithms optimise for the lowest misclassification error. This can result in probability distributions that underestimate or overestimate the true probability. This creates problems when we want to create a monotonic pattern but our model’s output is not monotonic.
* First, check for over/under-estimation using a reliability diagram. The  **model is well calibrated only if the calibration curve is very close to the bisector** since this would mean that the predicted probability is on average close to the theoretical probability.



* If not well calibrated, use a second model (called calibrator) like Isotonic regression, **that “corrects” them into real probabilities**. rescaling probabilities monotonically to ensure more orderly intervals that scale more proportionately with credit score.

[200 words]

**Q2.2 How would you account for the month on month variability in default rates?**

Seasonality

* Month-on-month variability in default rates is most likely due to external cycles or seasonalities e.g. business cycles, housing market cycles, seasonal needs e.g. end-of-year spending

Model Impact

* Some decreases in default rates could be due to the success of the credit risk model in eliminating high risk users.

Solution

* [Seasonality] Use indicator variables that correspond to that seasonality, such as timestamps, in our model-building. This will help to account for the variability. For example, if our logistic regression model shows that the betas are positive in certain months like January and February, then we can infer that users are more likely to repay at the start of the year.
* [Model Impact] Similarly, add indicator variables that correspond to the model’s impact e.g. number of high-risk users eliminated in the previous month. If this variable’s beta is positive, then we can infer that the model is impacting monthly default rates.

[153 words]

*Management is thinking of using individual product pricing on per user basis.*

**Q3.1 How would you design data collection to ensure that you can build a machine learning model on top of that?**

An individual product pricing for instalment repayment will vary three different variables (1) how much person X should pay per occasion (2) instalment periods (3) total amount to be paid. The pricing regime will have different packages of (1), (2), and (3) depending on the overall business strategy and demand from each person (or group).

To build an effective model, data collection will need to record, for every user, (A) which package(s) were offered (b) which package was selected (c) impact on repayment rate (d) impact on default.

With these, a new ML model can be built to predict the most likely package users will purchase. Additionally, our previous default prediction model can be enhanced with the additional feature of the users’ package.

[124 words]

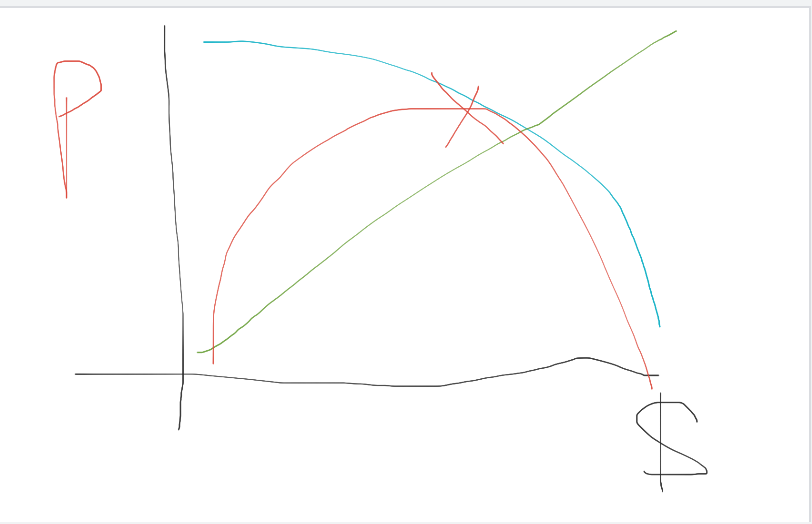
**Q3.2 Which model would you use and why to model the adoption probability for each user?**

A linear model e.g. logistic regression, where interpretability is higher, is best suited for this particular use case. Interpretability is important when designing a new product because we will need to isolate the marginal impact of changes in (A), (B), and (C) to adoption probability (and therefore, profitability). With the coefficients from the linear model’s output, we can e.g. observe the impact of increasing price by 1 dollar on adoption probability. Controlling for user’s characteristics, we can plot out the adoption probability curve for all possible price ranges, and find the optimal price point to maximise profitability.

[99 words]

**Q3.3 How would you account for and present the probability of adoption vs. profitability from pricing tradeoff?**

We can plot the expected **adoption probability** against the **profitability** at various price points. Multiplying the two gives us the **expected profitability**. At some point, the trade-off between the two is optimal. A very badly drawn graph is given below (with expected profitability scaled). The X-mark represents the optimal point that will most maximise profitability,



[55 words]