MGT 6203 Group Project Proposal

TEAM INFORMATION

Team #: 4

Team Members:

1. **Reiko Takizuka (edX: r-takizuka)** Reiko used to work as a researcher in a telecommunications company in Japan for over 25 years. She holds a bachelor's degree in engineering. She has finished CSE6040x and ISYE6501x and she hopes to enroll in the OMSA program next year to acquire additional skills in this field.
2. **Duc Nguyen Hong (edX: nhduc1993)** Duc holds a bachelor's degree in Finance & Banking and has established himself as a senior data analyst at an outsourcing company based in Vietnam. Duc is now preparing to embark on the next phase of his academic journey by enrolling in the OMSA program during the upcoming fall semester.
3. **Chandana Mudunuri (edX: ana_1009)** Chandana is a Customer Intelligence Analyst working for Racing and Wagering, WA in Australia. She has a bachelor's in electrical engineering and has accumulated more than ten years of experience in the data field, working in various roles such as data engineer, reporting engineer, and data analyst. She is eager to expand her knowledge in the field of data science and acquire additional skills to further enhance her expertise.
4. **Dayoon Sug (edX: DayoonSug)** Dayoon is IT Service Delivery Manager in Group Functions scope within South Korea at Merck Group. She has a bachelor's degree in business administration from Georgia Tech. This course is her first course studying data analytics.
5. **Michael Sullivan (edX: mikeys1)**. Mike is a Trade and Investment Analyst working for UK government. He is an economics graduate, has taken a Data Science bootcamp and will begin the OMSA this year.

OBJECTIVE/PROBLEM

Project Title: Bet on Better Forecasts: Leveraging Customer-Level Insights for Enhanced Financial Prediction for Racing and Wagering Western Australia (RWWA).

Background Info: Racing and Wagering Western Australia (RWWA) is the controlling authority for horse and greyhound in the state of Western Australia. It is a governmental body charged with maintaining the long-term sustainability of the racing industry in the region.

Part of the RWWA's remit is to control off-course betting activities for the racing industry, which they do through the Western Australia TAB[1], a state-owned company that runs over 300 betting retail outlets and an online betting platform known as "TABtouch". WA TAB allows customers to wager on racing (horse or greyhound) and other sports (both Australian and international). Each year customers place bets with WA TAB of around AUS $2bn, and WA TAB pays out around AUS $1.7bn to winning bets. Though large, these amounts make WA TAB a relatively small player in the context of the gambling industry.

[1] Totalisator Agency Board (commonly known as the TAB), which was a body created in 1961 as a result of the 1959 Royal Commission into off course betting.

Problem Statement: TABTouch's current approach to predicting future cash flow relies primarily on time series models built on aggregated data. While these models can be effective, we believe they might not capture all dynamics of TABTouch's customer behavior, nor do they allow for detailed exploration of evolving patterns among different customer segments. We believe that leveraging customer-level transaction data could allow us to provide more accurate cash flow predictions for the company and better understand which customer segments are driving business value.

Primary Research Question:
Can the use of customer-level wagering transaction data improve the accuracy and reliability of financial forecasts vs. aggregate time series methods?

Additional Questions:
- Can probabilistic Customer Lifetime Value (CLV) models and regression-based models predict the future value of individual customers with accuracy, and can they be integrated into our forecasting approach?
- What customer behaviors and characteristics are most useful for predicting future cash flows?
- How can customer segmentation enhance the effectiveness of our financial forecasting models and our individual level forecasting models?
- (Extension TBD): Can we accurately predict which customers might have gambling addiction issues in order to safeguard them?

Business Justification: Financial forecasting and predicting future cash flows is vitally important for TABTouch to make informed business decisions and plan strategically. They need to know how much cash they expect to take in to plan their overall operations and ensure the financial sustainability of the business. Knowing expected future cash flows helps the company decide how to spend money on staff, marketing, customer retention, technology improvements and many other areas. Without accurate cash flows predictions, business operations could be jeopardized. In our case this is doubly important because TAB and the RWWA support the entire racing industry in Western Australia – they need to be able to predict how much they can afford to spend on various programs to support and develop the industry.

DATASET/PLAN FOR DATA

Data Sources:

TAB Betting Data Our primary data source is (anonymized) internal wagering data from TAB: we have data that aggregates daily betting totals for each individual customer across 2021 and 2022. The link is to the full dataset (1GB+) and we have attached a smaller dataset for preview to this submission.

Labor Force Data: We will supplement our primary data with official Australian economic statistics on labor force participation and earnings.

Data Description: Primary dataset – TAB Betting data. Each row of this dataset represents the bets made by one customer on a given day. There is a unique customer identifying number, customer attributes and then a breakdown of the customers betting pattern for that day. Because of the sheer number of bets made on the platform, we are using data that has been aggregated daily rather than including one separate row for each bet made by each customer.

```
'data.frame':   143319 obs. of  16 variables:
 $ DATE_DIM            : chr  "2021-01-01" "2021-01-01" "2021-01-01" "2021-01-01" ...
 $ DAY_OF_WEEK         : chr  "Fri" "Fri" "Fri" "Fri" ...
 $ BET_ACCOUNT_NUM_HASH: num  13154 18379 559232 698904 762921 ...
 $ AGE                 : int  67 54 63 69 67 46 46 76 64 55 ...
 $ AGE_BAND            : chr  "65+" "45-54" "55-64" "65+" ...
 $ GENDER              : chr  "M" "M" "M" "M" ...
 $ TENURE_IN_DAYS      : int  11846 1884 2866 2100 4766 2307 714 10686 2663 5728 ...
 $ RESIDENTIAL_STATE   : chr  "WA" "WA" "WA" "WA" ...
 $ FOB_RACING_TURNOVER : num  37 40 NA NA NA NA 56 10 56 15 ...
 $ FOB_SPORT_TURNOVER  : num  NA NA NA NA NA NA 68 NA NA NA ...
 $ PARI_RACING_TURNOVER: num  1081 NA 12 1223.5 17.5 ...
 $ PARI_SPORT_TURNOVER : num  NA NA NA NA NA NA NA NA NA NA ...
 $ TOTAL_TURNOVER      : num  1118 40 12 1223.5 17.5 ...
 $ DIVIDENDS_PAID      : num  443.6 0 9.5 267.9 0 ...
 $ GROSS_MARGIN        : num  271.25 40 2.04 245.12 3.5 ...
 $ TICKETS             : int  288 1 5 40 5 11 12 1 46 7 ...
```

Key Variables: The dependent variables in our primary dataset for our primary research question are "GROSS_MARGIN" (the profit made on all bets placed by a customer on that day) and "TOTAL_TURNOVER" (the total monetary value of all bets placed by a customer on that day). The independent variables include demographic information such as "AGE", "GENDER", "RESIDENTIAL_STATE"; breakdowns of betting across racing vs. other sports (check for "RACING" vs. "SPORTS" in column names); breakdowns of betting different types of betting activities[2] (check for "FOB_" vs. "PARI_"), the total number of tickets purchased by the customer on that day ("TICKETS").

APPROACH/METHODOLOGY

Planned Approach: We plan to approach this project like a typical data science one with all the necessary processes.

Data Exploration & Preprocessing: We will clean the data (missing value imputation, duplicate detection, outlier detection) and explore seasonality, trends and non-stationary issues.

Feature Engineering & Customer Segmentation: We will engineer features from our customer-level data to use in segmentation and other modeling (such as frequency, average wage amounts, win/loss ratios, time since last active, etc). We also segment customers using methods like K-mean clustering and cohort analysis.

Model Development: We will create three different types of models.

(1) Time Series Models – as baselines we will build models (e.g. exponential smoothing, ARIMA) forecasting financial info on aggregated data.

(2) Probabilistic Customer Lifetime Value (CLV) Models – we will construct a probabilistic model (e.g. Pareto/NBD to predict frequency and Gamma-Gamma for values) to predict individualized customer level forecasts (which we will then aggregate).

(3) Regression Models – we'll construct regression models to predict individual customer values over various periods, making use of our engineered features. We will try various types including linear and non-linear regression, regression trees, random forests, etc.

Training, Optimization & Evaluation: We split our data temporally, use grid-search with cross-validation to tune models, and then assess performance on our final period test data. We will use metrics like RMSE and MAE to see if our customer-level models can beat our time series models.

[2] TAB allows users to participate in both Fixed Odds Betting and Parimutuel Betting

Refinement & Improvement: We will see whether developing new features or combining / blending our models in some way can lead to improved performance.

Anticipated Conclusions/Hypothesis: We expect that customer level data could generate more accurate financial forecasts. Our approach will allow us to test this on real unseen data. Whether the improvements will be enough to justify a much more complicated model is uncertain. We also have several other secondary hypotheses about our problem at this stage. We expect K-means clustering method will not be able to segment customers into meaningful groups with distinctive quality. We expect large seasonal patterns: turnover and gross margin follow a weekly cycle in which it rises on the weekends and dips on Monday; and sports/racing betting varies a lot in line with big calendar events. Betting frequency, win-loss ratio and total turnover of previous month are the most significant features to predict same metrics of next months. Probabilistic models might be very difficult to implement in our case because many of their assumptions (e.g. homogenous customers) don't seem to be true given what we have observed in the data.

What business decisions will be impacted / what are some benefits: By developing an enhanced financial forecasting model, TABTouch can gain a more accurate assessment of their budget allocation for marketing, technology upgrades, and customer retention initiatives. This improved understanding of customer segments and their impact on future value will enable TABTouch to plan their marketing and customer retention activities more effectively. For instance, with a 10% increase in forecasting accuracy, TABTouch can allocate an additional $500,000 to targeted marketing campaigns, allocate $200,000 towards technological advancements, and invest $300,000 in customer retention programs. These strategic investments driven by a comprehensive financial forecasting model will position TABTouch for increased revenue growth and improved customer loyalty.

PROJECT TIMELINE/PLANNING

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Phase	Description	Timeline
1	Team Formation	June 04
2	Finding a Project	June 05 -> June 07
	Data Preparation & Simple EDA	June 07 -> June 14
	Business Understanding & Feasibility	June 14 -> June 21
	Project Proposal Submission	June 21
	Deep Exploratory Data Analysis	June 21 -> June 28
	Project Proposal Presentation Video	July 02
3	Modeling & Validation and Project Progress Write-up	July 02 -> July 09
	Project Progress Report Submission	July 09
4	Final Report Write-up and GitHub Repo Finalization	July 09 -> July 20
	Final Report Submission	July 20
	Final Presentation Video	July 23
5	Peer Review: Out-of-Group Final Video Presentation	July 28
	Peer Review: Within-Group Performance Evaluation	July 28