Proposal: Predicting Uber Ride Cancellations via Machine Learning

*For the fulfillment project proposal of AT82.01 Computer Programming for Data Science and Artificial Intelligence course by Dr. Chantri Polprasert

1st Biki Nath Newa
Department of Data Sciences and Artificial Intelligence
Pathum Thani, Thailand
st126199@ait.ac.th

2nd Hein Min Htet
Department of Data Sciences and Artificial Intelligence
Pathum Thani, Thailand
st126459@ait.ac.th

***Abstract*—In our project proposal, we present a plan to forecast Uber ride cancellations using Machine Learning based on a dataset of ride requests from Kaggle. We will analyze features like pickup location, pickup distance and destination distance to build a robust classification model.**
***Index Terms*— Uber, Cancellation Prediction, Classification, Machine Learning,**

I. Introduction

A. *Ride-Hailing Services*

Ride-hailing services like Uber have become an integral part of modern urban transportation, not only providing a job opportunity for anyone that owns a vehicle and fulfills the requirements such a driving license, proof of residency and a background check. With the success of the very first ride-hailing app, more apps with a similar model has shown up in the market creating a strong contention for users. And as such, being able to minimize user churn is integral to success.

B. *Ride Cancellations*

A key operational challenge in this competitive ecosystem is the high rate of cancellations which can be initiated either by the rider or the driver. These cancellations disrupt the service's efficiency, leading to poor experience for users and lost income for drivers.

C. *Customers' Perspective on Ride Cancellation*

To discourage haphazard ride cancellations, the ride sharing apps have a policy of charging the users for cancellation. Even with the implementation of policies such as a grace period of 2 minutes where the user can cancel for free or not charge the user if the cancellation was done from the driver's side. Many users have expressed their dissatisfaction citing cases where they needed to cancel due to the driver not moving for minutes or having a chance of plans. As such, this can be identified as one potential cause of churn.

D. *Business Perspective for Ride Sharing*

From a business standpoint, ride cancellations directly impact revenue, user retention, and driver satisfaction. For a company like Uber, high cancellation rates can lead to:

- **Reduced Platform Trust:** Customers may become frustrated and switch to competing services.
- **Inefficient Driver Allocation:** Drivers waste time and fuel traveling to pickups that get cancelled.
- **Lost Revenue:** Every cancelled trip is a missed revenue opportunity.
- **Hidden Cost:** While the above impacts are obvious, it also leads to server electricity, cooling and computing resources which can easily stack up with the quantity of total rides being handled at any given time.

D. *Why Do a Cancellation Prediction Project?*

1. **Critical Point of Failure:** The efficiency of the ride sharing apps relies on successfully matching drivers with riders. Thus, cancellations represent a significant point of failure in this matching process.
2. **Improve Platform Efficiency:** By predicting cancellations, Uber can adjust its algorithm and pair drivers and users more intelligently to avoid pairing with high likelihood of cancellations.
3. **Enhance User Experience:** Reducing the frequency of cancellations and avoiding cancellation fees means riders get picked up more reliably, leading to less churn and higher customer satisfaction and loyalty.
4. **Increased Driver Earnings:** Minimizing time spent on cancelled trips also allows the drivers to complete more paid fares, increasing their earnings and reducing frustration.
5. **Reduced Environmental Impact:** Without a good predictive model, drivers waste fuel moving to a pickup location which is inevitably cancelled. It also reduces the burden on the communication-server infrastructure from bad matchups.

In summary, the cancellation prediction projects help minimize user dissatisfaction and churn while also maximizing profits. Furthermore, it also helps identify the various causes of cancellations and can inform future decisions made to address it.

II. PROBLEM STATEMENT

To develop a machine learning model that accurately predicts the likelihood of an Uber ride request being cancelled. The prediction will be based on the available features present in the dataset, including the time of the request, distance between driver and pickup location, distance to the destination and historical data trends.

III. RELATED WORKS

The challenge of optimizing ride-hailing services is a popular area of research. Many studies have focused on predicting demand or arrival times, which share methodologies applicable to predicting cancellations.

Aditi G and Ashish P [1] proposed an integrated approach to address both ride cancellations and fare estimation for services like Uber. They proposed leveraging predictive modeling and machine learning techniques on historical ride data. By developing models to forecast cancellation likelihood and a dynamic approach to fare estimation, they aimed to enhance user satisfaction, optimize driver allocation, and promote trust within the ridesharing ecosystem. Using Logistic Regression, Decision Tree and Gradient Boosting models, Gradient Boosting provided the best results with accuracy 0.97, precision 0.95, recall 1.0 and F1 score 0.97.

Xiaolei Wang, Wei Liu et al. [2] from China and Australia investigated customer order cancellations in coupled ride-sourcing and taxi markets. They proposed a behavioral model based on a two-month hourly-average dataset from Didi Chuxing. They concluded that customer cancellations waste driver efforts and lower the availability of supplies on the platform, and their model helps to understand the dynamics of customer choice and cancellation behavior when both ride-sourcing and traditional taxis are available.

S. Krishnaveni and M. Jaya [3] analyzed the problem of Uber trip cancellations. They proposed using time series modeling and a decision tree classifier. Using a dataset of Uber rides from Coimbatore, they concluded that the decision tree classifier could effectively predict trip cancellations and suggested that increasing cab availability during high-demand periods is a key factor in reducing them.

IV. DATASET

A. Description

The dataset for this project is sourced from Kaggle: **"Uber Ride Analytics Dashboard"** by Yash Dev Laddha. It contains a record of approximately 148,770 Uber ride requests during a specific period. The data is provided by a third-party analytics firm and is formatted as a single CSV file, making it readily accessible for analysis.

B. Features

The dataset contains the following features:

- Date and Time of Booking
- Booking ID, Customer ID
- Booking Status (Completed, Incomplete, Cancelled By Driver, Cancelled by Customer)
- Vehicle Type
- Pickup and Drop location
- Average Time to Customer
- Average Time from Pickup to Destination
- Cancelled By Customer, Customer Cancellation Reason
- Cancelled By Driver, Driver Cancellation Reason
- Booking Value
- Ride Distance
- Driver Rating
- Incomplete Ride, Incomplete Ride Reason
- Customer Rating
- Payment method

V. METHODOLOGY

This project will use Python's scikit-learn and pandas libraries for data analysis and modeling. The workflow will follow standard data science practices from data acquisition to deployment.

A. Data Acquisition

The dataset will be downloaded directly from the provided Kaggle URL.

B. Exploratory Data Analysis (EDA)

For this task, we will explore the data to see if there are any potential problems in the dataset such as outliers, mislabeled data and unwanted correlations between features.

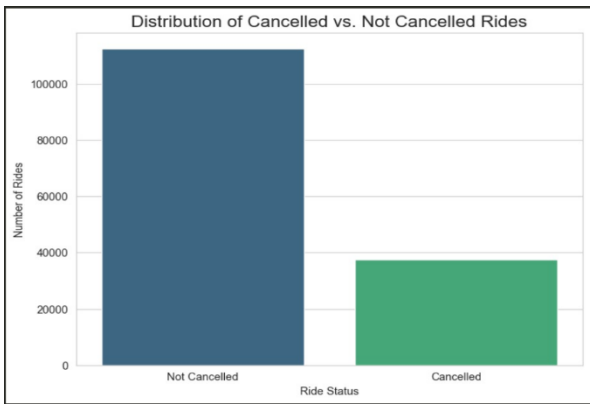

Fig. 1. Dataset Balance

```
df.isna().sum()
✓ 0.0s
Date                                  0
Time                                  0
Booking ID                            0
Booking Status                        0
Customer ID                           0
Vehicle Type                          0
Pickup Location                       0
Drop Location                         0
Avg VTAT                          10500
Avg CTAT                          48000
Cancelled Rides by Customer      139500
Reason for cancelling by Customer 139500
Cancelled Rides by Driver        123000
Driver Cancellation Reason       123000
Incomplete Rides                 141000
Incomplete Rides Reason          141000
Booking Value                     48000
Ride Distance                     48000
Driver Ratings                    57000
Customer Rating                   57000
Payment Method                    48000
dtype: int64
```

Fig. 2. Checking for NaN

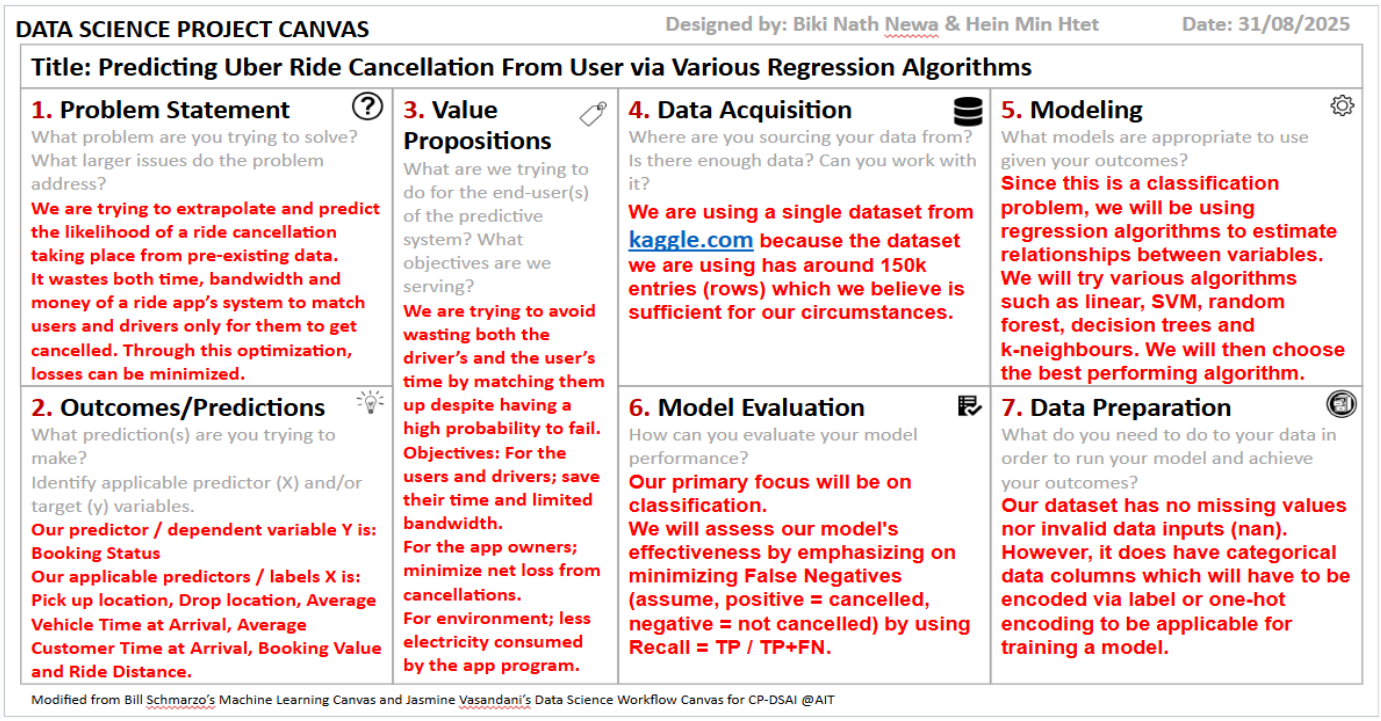

Fig. 3. Project Design Canvas

C. *Pre-processing*

After exploring what our data set looks like, we need to do pre-processing. The most pressing concerns are how we should handle the high amount of NaN, due to poor one hot encoding as well as the lack of a balanced dataset.

D. *Modelling*

Since this is a classification problem, we will experiment with several models starting with linear regression but we will probably need to settle for a more advanced model such as SVM or Random forest.

E. Training

Training will be done via the scikit-learn library .fit method. We will train several models with differing parameters to find the best performing model.

F. *Evaluation*

The models' performance will be evaluated on the test set using standard classification metrics using:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 * [(Precision * Recall) / (Precision + Recall)]
- R-Squared (R^2) = 1 - (Sum of Squared Residuals / Total Sum of Squares)

G. *Modelling*

The final, trained model will be deployed as a simple web application using Dash. The steps are as follows:

1. Save the trained scikit-learn model as a pickle file.
2. Create a Dash web app that loads the model.
3. Build a simple HTML interface where a user can input ride details (e.g., pickup point, time) and receive a cancellation prediction.

REFERENCES

[1] A. Gupta, A. Pipaliya, L. Jacob and K. Balagangadhar, "Predictive Modeling for Uber Ride Cancellation and Price Estimation: An Integrated Approach," *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, Pune, India, 2024, pp. 1-8, doi: 10.1109/TQCEBT59414.2024.10545287.

[2] Xiaolei Wang, Wei Liu, Hai Yang, Dan Wang, Jieping Ye, Customer behavioural modelling of order cancellation in coupled ride-sourcing and taxi markets, Transportation Research Procedia, Volume 38, 2019, Pages 853-873, ISSN 2352-1465, https://doi.org/10.1016/j.trpro.2019.05.044.

[3] Dr. S. Krishnaveni and M.Jaya Tharunigha, "PREDICT THE CANCELLATION OF TRIPS USING CLASSIFIERS AND TIME SERIES MODELING ", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.9, Issue 6, page no.235-239, June-2022, http://www.jetir.org/papers/JETIRFN06039.pdf