Predicting Uber Ride Cancellations via Machine Learning

by

ST126199 - Biki Nath Newa
ST126459 - Hein Min Htet

ASIAN INSTITUTE OF TECHNOLOGY SCHOOL OF ENGINEERING AND TECHNOLOGY DEPARTMENT OF INFORMATION AND COMMUNICATIONS TECHNOLOGY

Course	: Computer Programming for Data Science and Artificial Intelligence
Submit to	: Dr. Chantri Polprasert
Submitted by (**When September Ends**)	: Mr. Hein Min Htet (ST126459) Mr. Biki Nath Newa (ST126199)
Date of Submission	: 30th November 2025

Asian Institute of Technology

School of Engineering and Technology

Thailand

Aug 2025

AUTHOR’S DECLARATION

Hein Min Htet, Biki Nath Newa, declared that the research work carried out for this thesis was in accordance with the regulations of the Asian Institute of Technology. The work presented in it is our own and has been generated by us as the result of our own original research, and if external sources were used, such sources have been cited. It is original and has not been submitted to any other institution to obtain another degree or qualification. This is a true copy of the thesis, including final revisions.

Date: 30th Nov, 2025

Abstract

This project addresses the critical operational challenge of forecasting Uber ride cancellations, a frequent occurrence that results in customer churn, significant revenue loss and logistical inefficiency for ride-sharing platforms. We propose a robust Machine Learning framework to accurately predict whether a solicited uber ride request will ultimately be canceled by the rider.

Our methodology involves developing and comparing several classification models (e.g., Logistic Regression, Random Forest, Gradient Boosting) trained on a publicly available, real-world dataset of Uber ride requests sourced from Kaggle. We meticulously analyzed and engineered key predictive features, including pickup location attributes, pickup and destination distance, time-of-day variables, weather factors, and driver-and-rider historical behavior metrics.

The primary objective was to select the model demonstrating the highest predictive accuracy to create ride hailing corporations with a tool for proactive intervention, ultimately minimizing cancellations and optimizing fleet management.

Index Terms: Uber, Ride Cancellation Prediction, Classification, Machine Learning, Predictive Modeling, Random Forest, XGBoost.

CHAPTER 1

INTRODUCTION

1.1 Modern Urban Transportation Operational Challenges (Uber)

Ride-hailing services, exemplified by Uber, are a cornerstone of modern urban transit, operating within a highly competitive market where operational efficiency and user retention are paramount.

A critical operational challenge within this system is the presence of ride cancellations, initiated by either the driver or the rider, which represents a significant point of failure in the platform's core matching process.

1.2 Impact on Operations

These cancellations severely impact user experience and often lead to frustration and user churn, despite the recent relaxing of existing cancellation fee policies. Simultaneously, they cause substantial business inefficiencies, including:

- Lost revenue
- Wasted driver fuel and time
- Unnecessary demands on server infrastructure

1.3 Objective of the Prediction Project

Therefore, the objective of this cancellation prediction project is to leverage Machine Learning to accurately forecast potential trip failures, enabling the corporation to proactively adjust their matching algorithms or enact churn management measures.

This predictive capability is essential for:

- Minimizing user dissatisfaction.
- Maximizing driver profitability through an increased number of completed fares.
- Enhancing overall platform reliability.
- Ensuring the intelligent allocation of resources.

1.4 Related Work

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.

Xiaolei Wang, Wei Liu et al. [2] - from China and Australia investigated customer order cancellations in coupled ridesourcing and taxi markets.

S. Krishnaveni and M. Jaya [3] - analyzed the problem of Uber trip cancellations.

1.5 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.

1.6 Methodology

- Data Filtering/Cleaning: Filtering the dataset to focus only on completed or specifically cancelled rides by user.
- Feature Engineering (Time-based): Extracting features such as hour, day, month, and day_of_month from the booking timestamp.
- Imputation (Group Mean): Filling missing Ride Distance and Booking Value data for cancelled rides using the calculated average values for the corresponding Pickup Location and Drop Location groups.
- Imputation (Forward Fill): Filling missing Customer Rating values based on a customer's most recent prior rating, a method that uses sequential data to prevent data leakage.
- Feature Engineering (Rating Binning): Converting numerical driver and customer ratings into ordinal categories (Poor, Average, Good, Not Rated).
- Feature Engineering (Historical Aggregation): Creating lagged features like historical_patience_vtat and historical_customer_avg_rating using time-series/group-based aggregation to measure historical behavior without data leakage.

- External Data Integration / Time-Series Join: Fetching supplementary weather data via an external API and merging it with the ride data using a time-series join to align the ride time with the most recent hourly weather reading.
- Feature Engineering (WMO Code Mapping): Mapping complex numeric WMO weather codes into simplified categorical descriptions (Clear, Rain, Drizzle, etc.).
- Feature Engineering (Location Zoning): Creating broad geographical categories (pickup_zone, drop_zone) from specific location names to generalize spatial attributes.
- Feature Transformation (One-Hot Encoding): Converting all remaining categorical variables (including vehicle type, day, and engineered features) into numerical columns suitable for machine learning using pd.get_dummies.

1.7 Datasets

1. Uber Data Analytics Dashboard

Url: https://www.kaggle.com/datasets/yashdevladdha/uber-ride-analytics-dashboard

This dataset contains 21 features excluding the index. They are:
Date, Time, Booking ID, Booking Status, Customer ID, Vehicle Type, Pickup Location, Drop Location, Avg Vehicle Time at Arrival (Driver to Customer), Avg Customer Time at Arrival (Customer to Destination), Cancelled Rides by Customer, Reason for cancelling by Customer, Cancelled Rides by Driver, Driver Cancellation Reason, Incomplete Rides, Incomplete Rides Reason, Booking Value, Ride Distance, Driver Ratings, Customer Rating, Payment Method.

Our main reason for settling on this dataset was that it was the only dataset that we could find which recorded information about rides which were cancelled.

2. Historical Weather Data From Meteo API

Url: https://open-meteo.com/en/docs/historical-weather-api

We made use of historical data from publicly available Open-Meteo weather API to further augment the original dataset and better assist the model in modelling all the relationships that lead to the cancellation of a ride.

Other Considered Previously Considered but Ultimately Disregarded:

The project relies mainly on a single uber dataset obtained from Kaggle.com. Unfortunately, we were unable to find any other suitable datasets in order to meet the criteria of using at least three datasets.

The datasets mentioned below could not meet the quality and availability of appropriate information provided by the first dataset Uber Data Analytics Dashboard by Yash Dev Laddha.

Furthermore, all these uber datasets provided below do not contain any rows where rides were cancelled.

https://www.kaggle.com/datasets/yasserh/uber-fares-dataset
https://github.com/Geo-y20/Uber-Rides-Data-Analysis/blob/main/UberDataset.csv
https://catalog.data.gov/dataset/2023-yellow-taxi-trip-data
https://developer.uber.com/docs/businesses/data-automation/data-download
https://www.kaggle.com/datasets/abhi231092/uber-rides-data-bw-city-and-airport

CHAPTER 2

EVALUATION

2.1 Primary Evaluation Metrics

The project uses 4 primary evaluation metrics for its classification models:

Metric	Reason
1. Precision	Precision is effectively how many of the predicted cancellations were actually cancelled. We decided to prioritize precision as it minimizes false interventions. Since corporate intervention strategy to staunch cancellation will consist of providing discounts for waiting longer, offering it in cases where the model isn’t as sure will be financially costly for the client.
2. AUC-ROC Curve	The AUC-ROC measures the model's ability to rank positive predictions (cancellations) higher than negative predictions (completed rides) across all possible thresholds. This allows us to assess the overall quality of the prediction probabilities.
3. Precision-Recall Curve	The PR Curve helps by ignoring the massive number of "safe" rides and focuses on how well the model handles the cancellations. The curve represents a sliding scale of thresholds. Every point on the line corresponds to a probability cutoff (e.g., 0.2, 0.5, 0.8) and one of the operating points is chosen which makes the most sense for the problem type (precision vs recall).

4. Cumulative Accuracy Profile Curve	The CAP Curve is a powerful metric that translates the model's abstract accurate-ness into a concrete "Return on Investment" (ROI) which allows us to visualize our returns better in regards to the amount of money we are willing to spend to placate a frustrated user.

2.2 Model Justification

We primarily worked on two machine learning model families, Random Forest and XGBoost. They are the industry standard for tabular data of sizes where we exceed a few hundred rows. Similarly Gradient Boosting Machines like XGBoost are the undisputed state of the art for these types of problems and often outperform Neural Networks (Deep Learning) on this type of structured data.

2.3 Training Strategy

When we carried out Pearson correlation and mapped it onto a heatmap, it was self-evident that the relationship between the features and the target were not linearly identifiable. Thus, we could not risk omitting any features. The tree models were able to automatically identify and capture the appropriate feature importance plots for this non-linear relationship.

Then we evaluated the model classification reports of each of the models while prioritizing precision. After settling on the most suitable model, we carried out grid search to attempt to find the most suitable parameters for the particular model.

2.3.1 Trained Models - Classification Report

Model 1 & 2: Random Forest Without SMOTE &

Stratified Random Forest Without SMOTE

	precision	recall	f1-score	support
0	0.98	1.00	0.99	18600
1	1.00	0.32	0.48	447
accuracy			0.98	19047
macro avg	0.99	0.66	0.74	19047
weighted avg	0.98	0.98	0.98	19047

Model 3 & 4: Random Forest With SMOTE &

Stratified KFold Random Forest With SMOTE

	precision	recall	f1-score	support
0	0.98	1.00	0.99	18600
1	1.00	0.32	0.48	447
accuracy			0.98	19047
macro avg	0.99	0.66	0.74	19047
weighted avg	0.98	0.98	0.98	19047

Model 5: Stratified KFold XGBoost (No SMOTE)

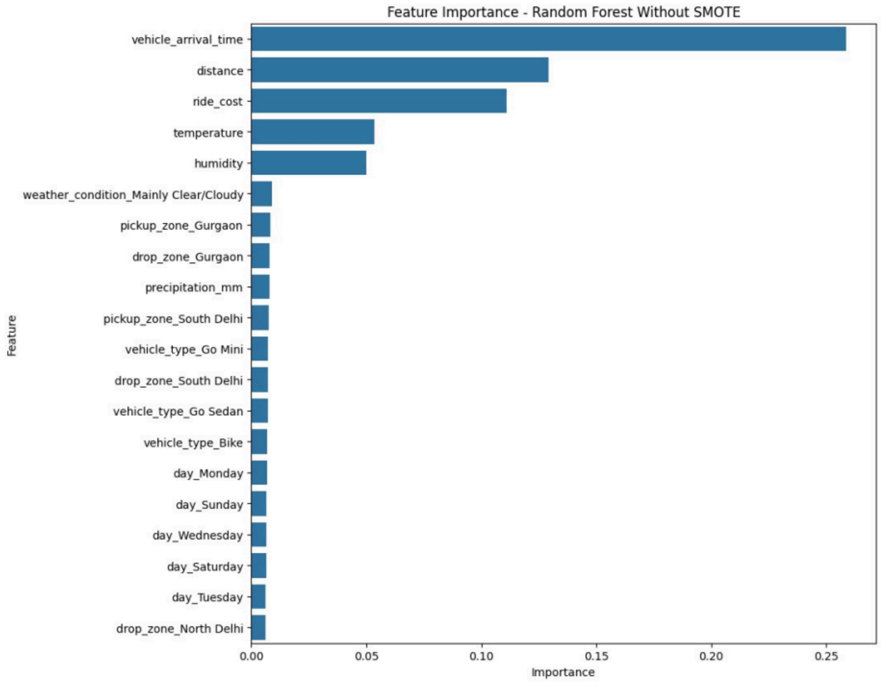
	precision	recall	f1-score	support
0	0.98	0.99	0.99	18600
1	0.47	0.34	0.39	447
accuracy			0.98	19047
macro avg	0.73	0.66	0.69	19047
weighted avg	0.97	0.98	0.97	19047

Selected Model: Stratified Random Forest

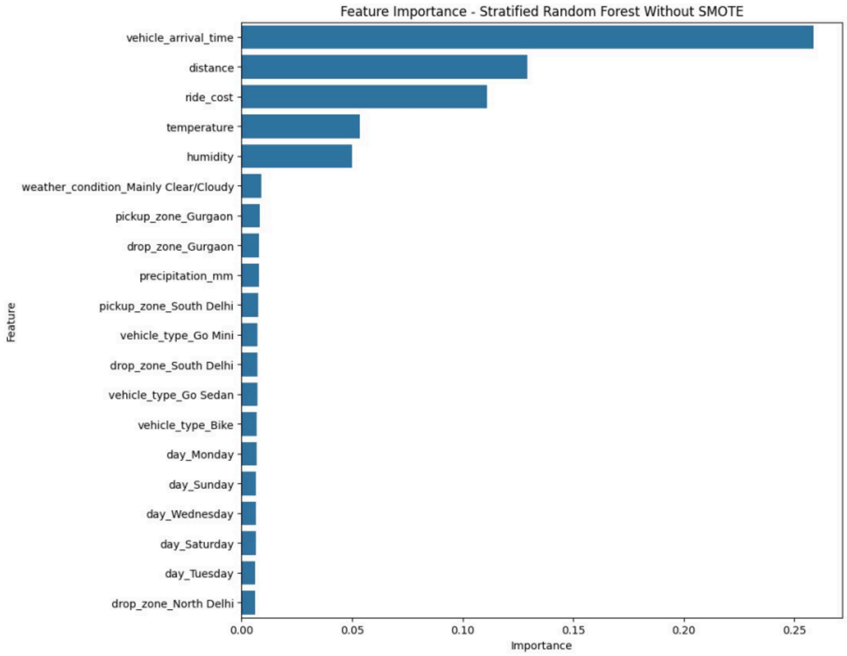
Metric	Class 0 (Not Cancelled)	Class 1 (Cancelled)	Overall Accuracy	Macro Avg	Weighted Avg
Precision	0.98	1.00		0.99	0.98
Recall	1.00	0.32		0.66	0.98
F1-score	0.99	0.48	0.98	0.74	0.98
Support	18,600	447	19,047	19,047	19,047

2.3.2 Feature Importance

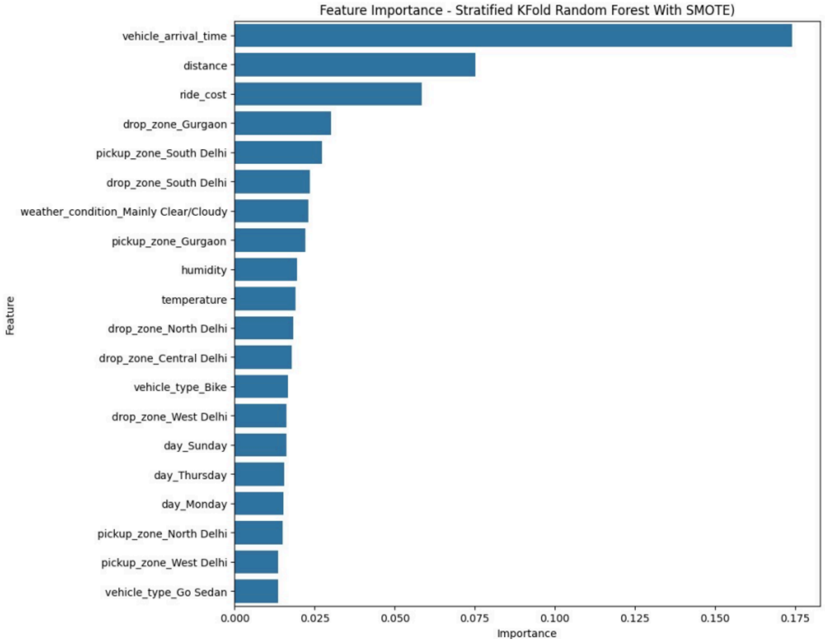
Model 1 - Random Forest Without SMOTE



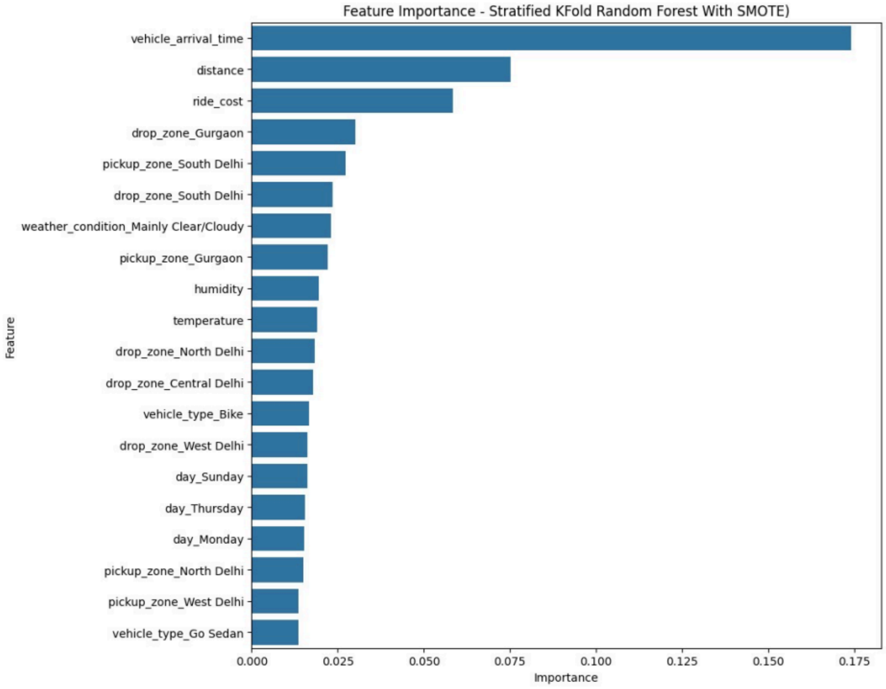
Model 2 - Stratified Random Forest Without SMOTE



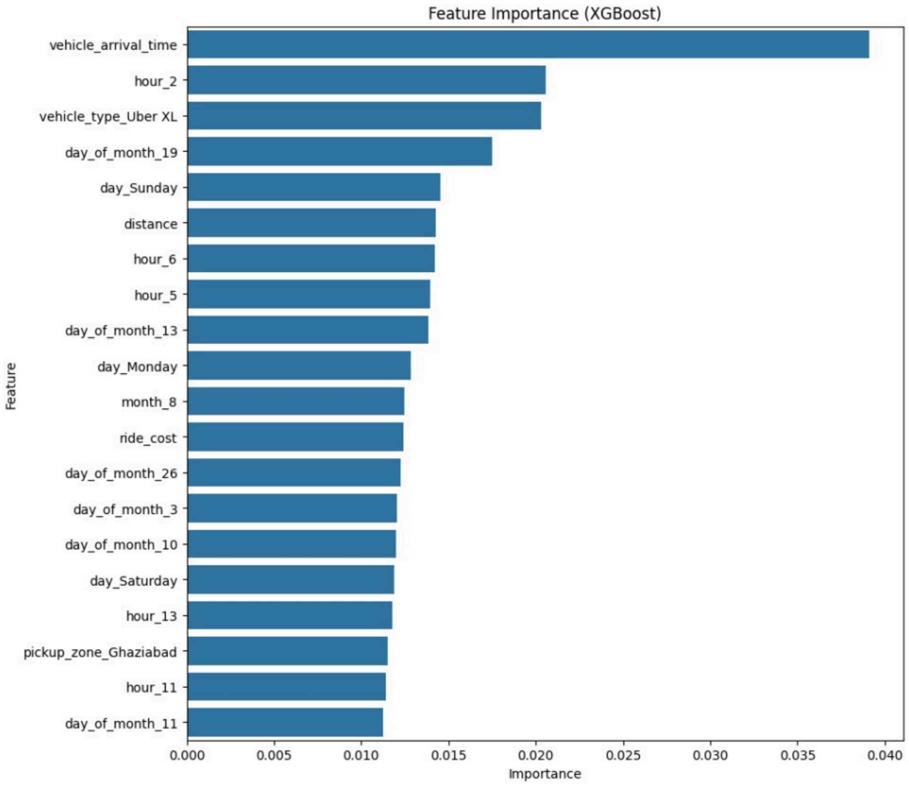
Model 3 - Random Forest With SMOTE



Model 4- Stratified KFold Random Forest With SMOTE



Model 5: Stratified KFold XGBoost (Without SMOTE)



2.3.3 Summary of Comparison

This project's success is measured by high overall accuracy of the model, but by its ability to reliably detect the minority class (cancellations). As we trained our models and looked at their results, we noticed that both the Random Forest and XGBoost models converged around roughly similar performances (Precision: ~1.00, Recall: ~0.34). Even when we added SMOTE, weights, and weather data, the recall refused to budge significantly.

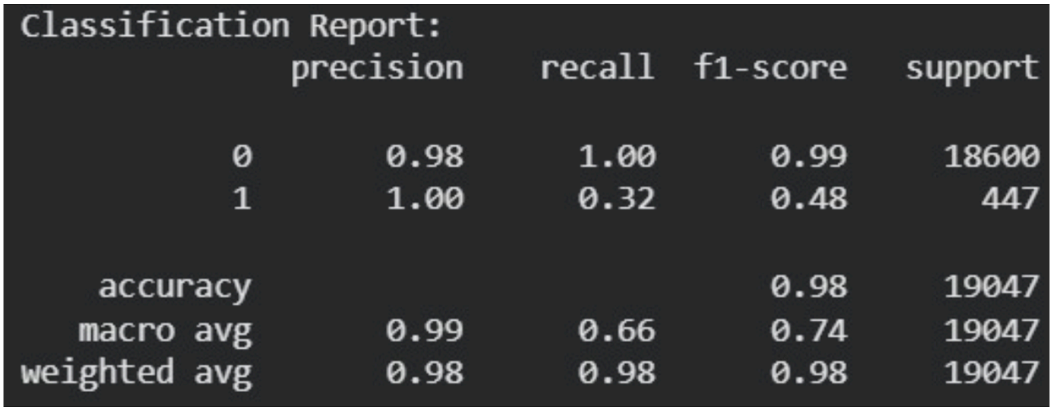
This strongly suggested that the cancellations that our models were missing looked mathematically identical to completed rides leading to an Aleatoric Uncertainty problem. No model can predict these because the data isn't there. We believe that the unpredictable cancellations were likely caused by the following:

1. **Driver Behavior:** Right now our strongest indicator of a cancellation is the waiting time. However, in cases where the ride was cancelled very early, it is possible that rather than the user being impatient, the driver might have been driving in the wrong direction which prompted the user to cancel. However, since the dataset lacks GPS traces and attempts at procuring traffic related data proved ineffective, this cannot be modelled.
2. **User Context:** Despite our attempt to reduce random cancellations by omitting data where the drivers did the cancellation or the cancellations occurred for some other reasons such as the vehicle being overcrowded or AC not working. It is possible that some of the users either imputed the wrong reason by accident or deliberately to avoid a penalty.
3. **App Glitches:** Lastly, we may also have circumstances where the ride hailing app might have glitched and dropped the user or the user may have lost connection causing the cancellation.

   The reason that makes our dataset unique is the fact that it still contains rows of incomplete and cancelled rides. However, the same property could have been working against us through this possibility.

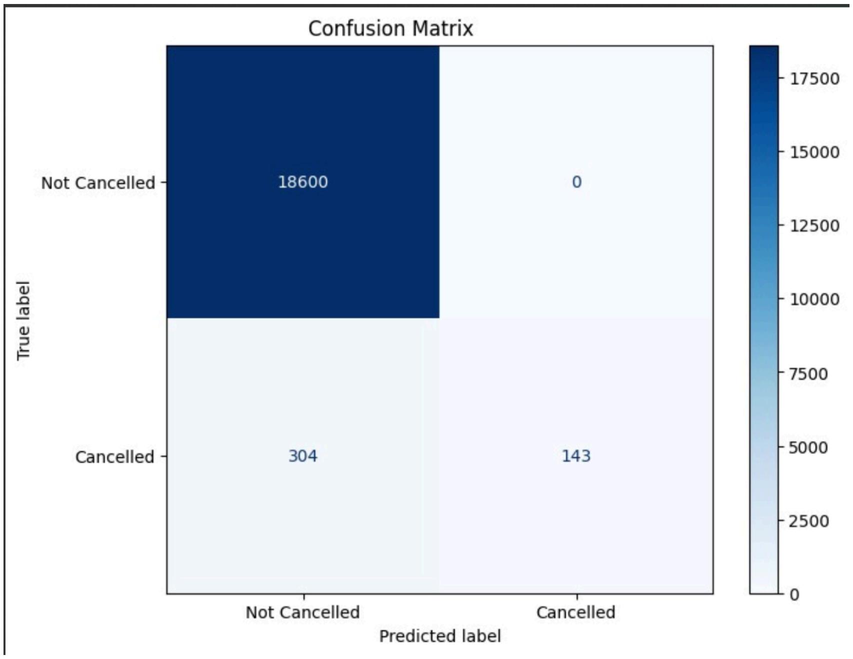
2.4 Final Model Performance Results

Classification Report

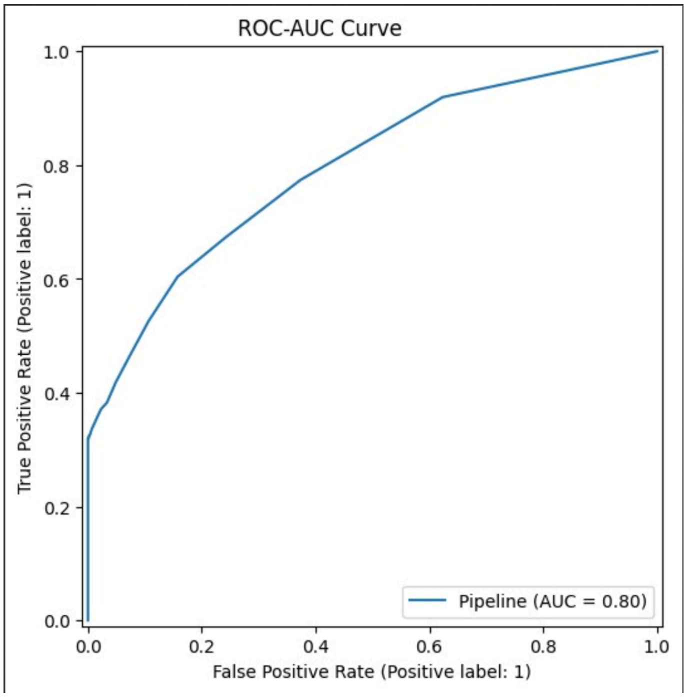


```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     18600
           1       1.00      0.32      0.48       447

    accuracy                           0.98     19047
   macro avg       0.99      0.66      0.74     19047
weighted avg       0.98      0.98      0.98     19047
```

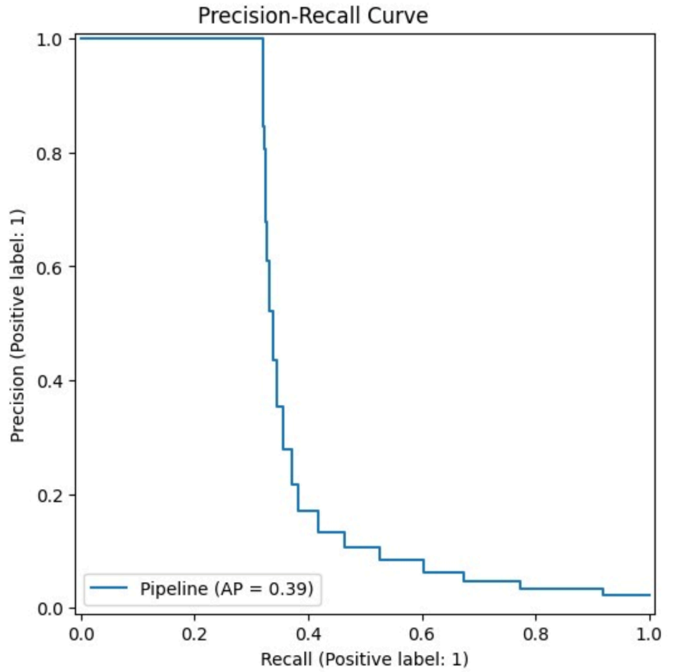
Confusion matrix



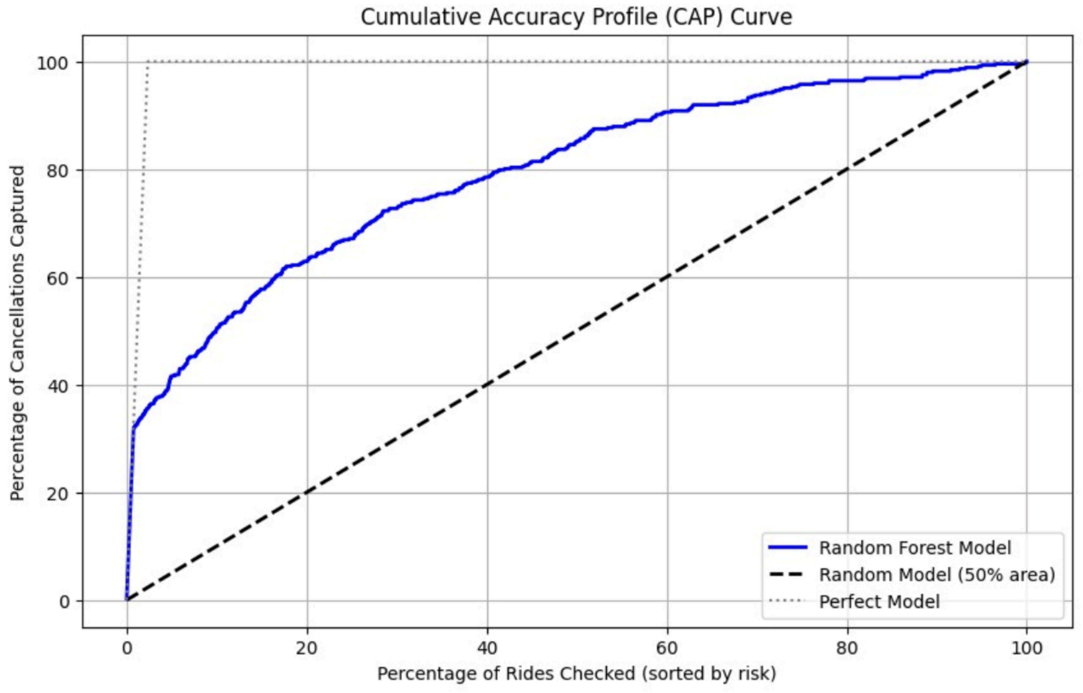
ROC-AUC Curve



Precision-Recall Curve



CAP Curve



CHAPTER 3

DISCUSSION

3 Discussion

3.1 Methodology and Model Robustness

The project employed a robust Machine Learning pipeline approach to predict ride cancellations. This methodology ensures that the data preparation steps including scaling numerical features and one-hot encoding categorical features are performed solely on the training data, preventing data leakage into the test set.

The core model used was a Random Forest Classifier. Critically, this model was initialized with the parameter class_weight='balanced'. This setting is essential for addressing the problem's inherent class imbalance where completed rides significantly outnumber cancellations and force the model to pay closer attention to the rare, high-cost cancellation events.

The features engineered for the model included core trip metrics (ride_cost, distance, vehicle_arrival_time), temporal variables (hour, day), spatial features (pickup_zone, drop_zone), and behavioral metrics (historical_customer_rating_binned).

3.2. Evaluation and Interpretation of Results

The model was evaluated on a test set containing 19,047 rides, of which only 447 were actual cancellations.

Metric	Class 1 (Cancelled) Result	Operational Interpretation

Precision	1.00	Intervention Reliability: For every ride the model predicted would be cancelled, it was correct. This translates to zero False Positives, meaning that any proactive intervention suggested by the model (e.g., providing an incentive, re-matching a driver) is guaranteed to be necessary and is never a wasted cost.
Recall	0.32	Coverage of Risk: The model correctly identified 32% of all actual cancellations. This indicates that while the model's predictions are highly reliable, it still misses $68 of the total potential revenue loss and logistical failures caused by cancellations.
F1-score	0.48	Balanced Performance: The F1-score provides a robust measure of performance on the minority class, combining the trade-off between Precision and Recall. The value of $0.48 shows moderate success in predicting the difficult-to-detect cancellation class.
Overall Accuracy	0.98	General Performance: The high overall accuracy of $98 is misleading due to the severe class imbalance, confirming why the F1-score is the more relevant business metric for this project.

3.3 Key Predictive Factors (Feature Importance)

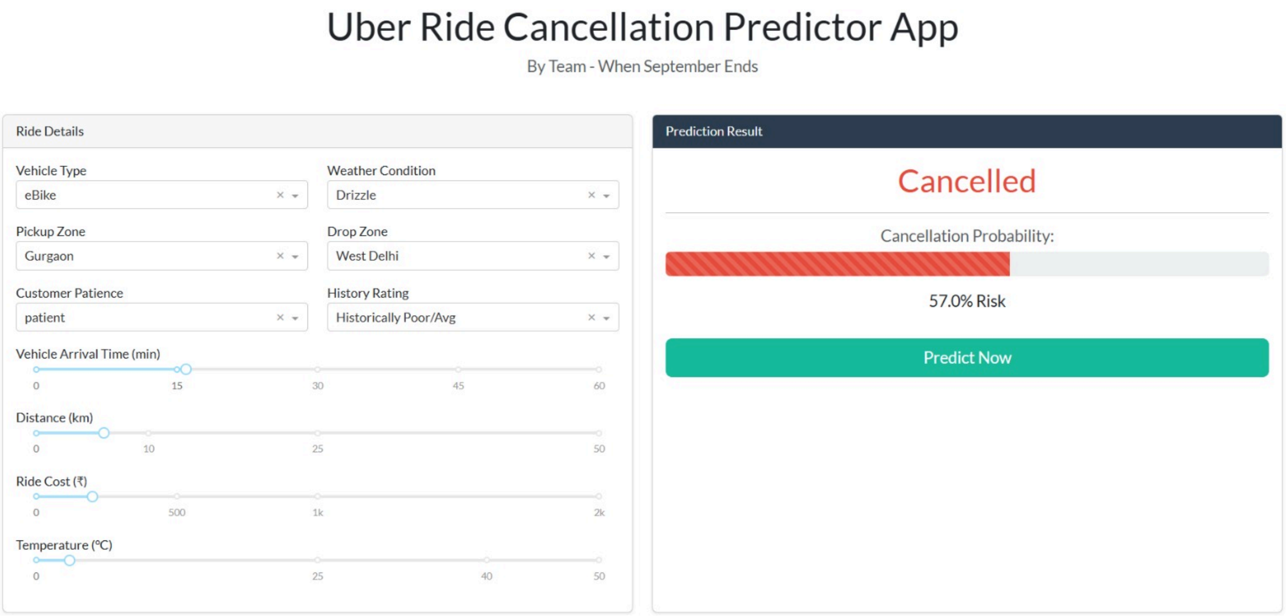
Analysis of feature importance reveals that the model relies heavily on core transactional and contextual data:

- Core Trip Metrics consistently rank as the most influential factors, confirming that the economic and convenience aspects of the trip are the primary drivers of cancellation risk.
- Geographical and Temporal Data are also critical. This suggests that cancellation risk is highly dependent on local supply-demand conditions and time-of-day traffic patterns.

3.4 Model Deployment

After the successful training of our model, we adopted a micro-service architecture where the machine learning model is decoupled from the training environment and served via a lightweight web application.

Instead of saving just the model, we serialized the entire Scikit-Learn Pipeline using .joblib which saves the scalers as a single object. This guarantees that raw input data from the web app is preprocessed exactly the same way as the training data, eliminating training-serving skew.



Uber Ride Cancellation Predictor App

By Team - When September Ends

Ride Details

Vehicle Type: eBike

Weather Condition: Drizzle

Pickup Zone: Gurgaon

Drop Zone: West Delhi

Customer Patience: patient

History Rating: Historically Poor/Avg

Vehicle Arrival Time (min)

Distance (km)

Ride Cost (₹)

Temperature (°C)

Prediction Result

Cancelled

Cancellation Probability:

57.0% Risk

To make the model accessible to non-technical stakeholders (e.g., Uber Operations Managers), we developed an interactive web dashboard using Plotly Dash. After the creation of an interactive Dash application our model was deployed on an AWS EC2 Amazon server.

By encapsulating our complex model within a user-friendly Dash interface and deploying it on cloud infrastructure, we transformed a theoretical analysis into an actionable tool. This allows the business to proactively identify high-risk rides in real-time and intervene before a cancellation occurs.

CHAPTER 4

CONCLUSION

4.1 Conclusion

The project successfully achieved its objective of developing a robust Machine Learning framework to predict Uber ride cancellations, providing a powerful tool for proactive operational intervention.

Project Achievements

- Robust Methodology: By implementing a Scikit-learn Pipeline, the project ensured an industry-standard approach that integrated preprocessing (scaling and encoding) with model training, completely preventing data leakage and guaranteeing that performance metrics reflect real-world deployment accuracy.
- Effective Imbalance Handling: The inherent class imbalance of the dataset was addressed by training a Random Forest Classifier with class_weight='balanced'. Evaluation utilized the F1-score, Precision, and Recall metrics, as well as the ROC and Precision-Recall Curves.

The current Random Forest model serves as an excellent baseline for a production-ready intervention system due to its (1.0) precision on the cancellation class. The model is exceptionally reliable. When it predicts a cancellation, it is virtually never wrong and thus a business can confidently implement this model knowing that every alert it generates will directly prevent a cancellation, and maximize the return on investment for any proactive measure with least loss.

4.2 Operational Impact

The final model's performance demonstrates a high degree of immediate operational utility:

- **Zero Waste Intervention:** The model achieved an exceptional precision of $1.00 (100%) for the 'Cancelled' class. This translates directly to zero False Positives, ensuring that any automated intervention (e.g., driver incentives, re-matching algorithms) triggered by the model will be applied only to a genuinely at-risk ride. This guarantees that operational resources are never wasted.
- **Targeted Risk Coverage:** The model registered an F1-score of 0.48 on the cancellation class, indicating moderate success in classifying the difficult-to-predict minority events.

4.3 Future Work and Optimization

While the perfect Precision is a massive initial success, the model's Recall of $0.32 reveals the primary area for improvement. The current model still misses $68 of actual cancellations, representing significant uncaptured revenue and customer churn risk.

Future efforts should focus on:

1. **Threshold Calibration:** Implementing threshold optimization techniques to find a new operating point that strategically sacrifices a small amount of Precision (e.g., accepting a Precision of 0.90 or 0.95) to achieve a much higher Recall, thereby maximizing the overall profit by mitigating a greater number of high-cost cancellations.
2. **Cost-Sensitive Tuning:** Fine-tuning the model using business cost matrices that quantify the specific financial penalties for false positives (wasted intervention cost) versus false negatives (lost revenue) to achieve true profit optimization.
3. **Threshold Optimization:** The perfect precision suggests the model's default classification threshold is too strict. A business-driven analysis should be performed to tune the probability threshold to achieve a more acceptable balance, such as increasing Recall to 0.60 or 0.70 while maintaining a Precision above 0.90.
4. **Advanced Cost-Sensitive Learning:** Directly incorporate the business cost of a False Negative (missed cancellation/lost revenue) versus a False Positive (wasted intervention cost) to train a model that optimizes profit rather than just a statistical metric.
5. **Model Exploration:** Implement more powerful classification algorithms, such as CatBoost or LightGBM, which are often better at modeling complex, non-linear patterns and handling imbalanced data than Random Forest.
   Our dataset is heavy on categorical features, and our use of One-Hot Encoding which were necessary inadvertently creates a sparse matrix and can sometimes lose information. CatBoost handles categorical columns natively without One-Hot Encoding and often yields slightly better results with less tuning.
6. **Attempt Isolation Forest:** Also known as Anomaly Detection, since the cancellations are rare in our dataset, instead of trying to classify "0 vs 1", we could spin the problem around with an Isolation Forest which tries to learn what "Normal" looks like and flags anything "Abnormal" as a cancellation. Although it is a wildcard approach, it can sometimes work better when the class imbalance is extreme.
7. **Incorporate Traffic Data:** According to EDA, contrary to conventional beliefs, the patience of the user increases rather than decreases during high-traffic rush hours. If we can find and incorporate data, we would be able to better model the relationship between driver behavior and early customer cancellations.

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, https://doi.org/10.1016/j.trpro.2019.05.044

[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