Background of work

High ride cancellations by riders or drivers reduce efficiency, disrupt matching, and cause poor user experience with income loss for drivers.

Why we want to do?

Ride cancellations are a critical failure in ride-hailing services, reducing efficiency, trust, and satisfaction while causing driver income loss, wasted time, customer churn and business revenue decline with hidden costs.

Proposal: Predicting Uber Ride Cancellations via Machine Learning

Biki Nath Newa, Hein Min Htet

Business View: Cancellations not only affect revenue but also raise operational costs, lower platform credibility, and increase customer churn.

Possible Impact

Reduced Platform Trust

Users may switch to competitors.

Inefficient Driver Allocation

Drivers waste time and fuel on cancelled rides.

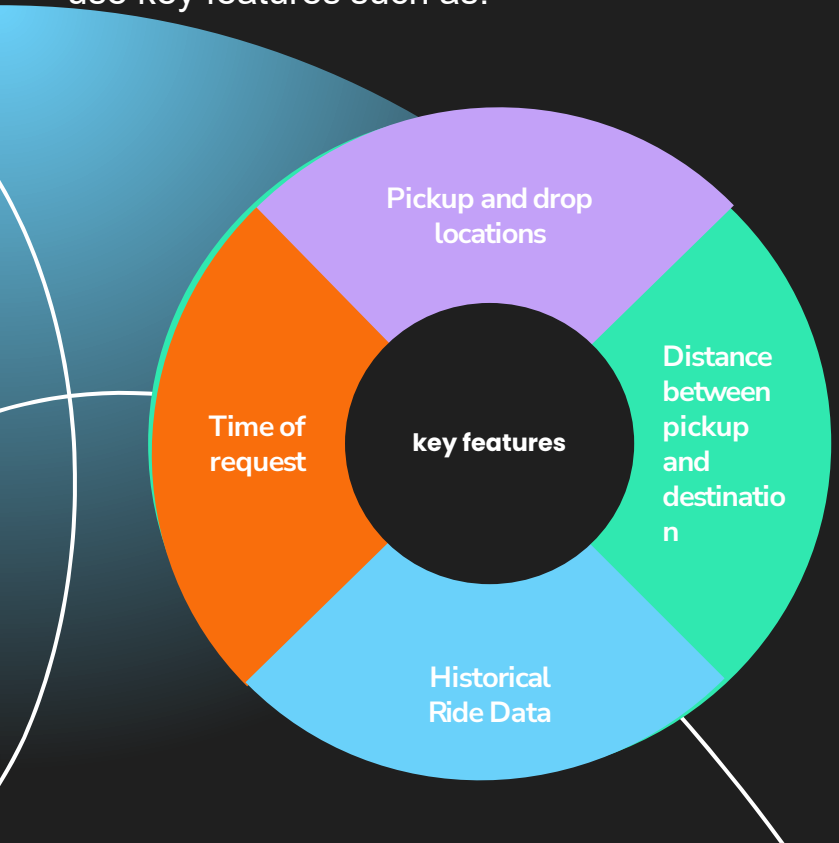
Hidden Costs

Extra burden on servers, computing, and energy.

Overall Risk

Weakens long-term customer loyalty and business growth.

We aim to develop a **machine learning model** that accurately predicts the likelihood of a ride being cancelled. The model will use key features such as:



Problem Statement

The high rate of ride cancellations in services like Uber is a critical challenge. These cancellations—initiated by either riders or drivers—reduce efficiency, disrupt the driver-rider matching process, and cause dissatisfaction, lost income, and hidden costs.

Related Works

Predictive Modeling for Uber Ride Cancellation and Price Estimation

Object

This study's goal by Aditi G and Ashish P is to address two main challenges in ride-sharing: ride cancellations and fare estimation. The researchers used a single, integrated approach instead of two separate ones.

Scope

The project used machine learning on historical ride data to build models for both **cancellation prediction** and **fare estimation**. They tested Logistic Regression, Decision Tree, and Gradient Boosting, finding Gradient Boosting to be the best with accuracy 0.97, precision 0.95, recall 1.0 and F1 score 0.97.

Related Works

Predict The Cancellation of Trips Using Classifiers And TIme Series Modelling

Object

The main objective of this study by S. Krishnaveni and M. Jaya was to **analyze the problem of Uber trip cancellations** and to build a model that could predict them.

Scope

Using a dataset from Coimbatore, the study applied **time series modeling** and a **decision tree classifier** to predict Uber cancellations. They found the decision tree classifier effective with 0.968 accuracy and recommended increasing cab availability during peak hours.

Dataset of Uber Ride Analytics Dashboard

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

Uber Ride Analytics Dataset 2024

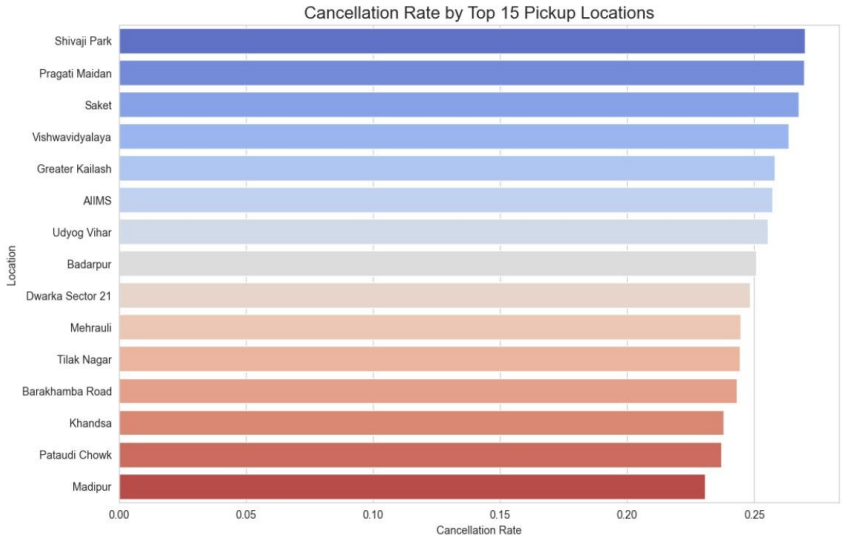
- **Total Bookings**: 148.77K rides
- **Success Rate**: 65.96% (93K completed rides)
- **Cancellation Rate**: 25% (37.43K cancelled bookings)
- **Customer Cancellations**: 19.15% (27K rides)
- **Driver Cancellations**: 7.45% (10.5K rides)

Data Acquisition

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

EDA- Exploratory Data Analysis

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.



Pre-processing

Modelling

Training

Evaluation

Deployment

EDA- Exploratory Data Analysis

Pre-processing

After exploring what our data set looks like, we need to do preprocessing. 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.

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.

Training

Training will be done by the scikit-learn .fit method. We trained the best model with different parameters to find the best performing model and parameters.

Evaluation

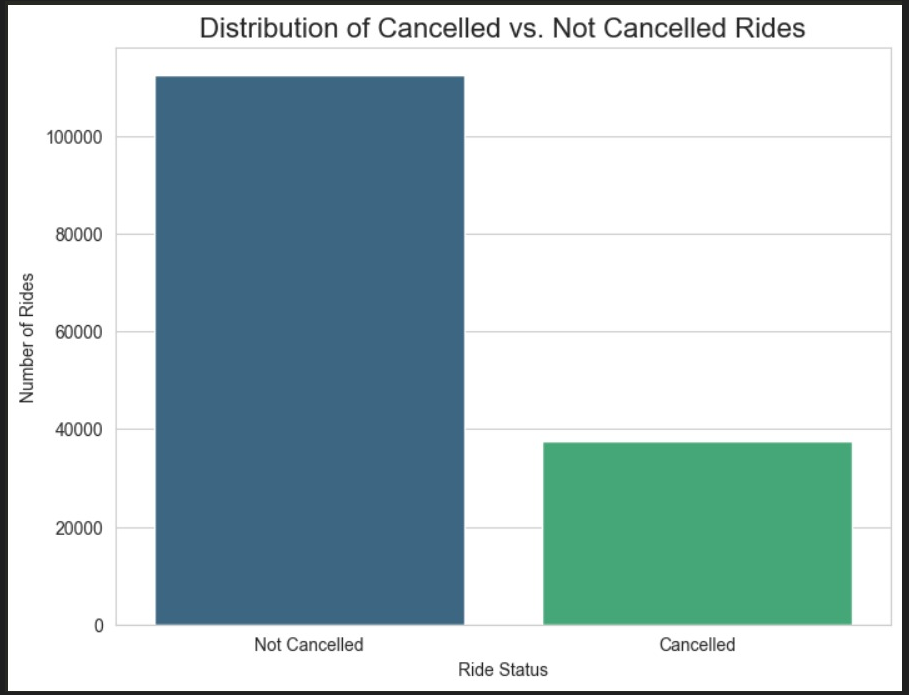
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 * [(Precision * Recall) / (Precision + Recall)]
- R-Squared (R^2) = 1 - (Sum of Squared Residuals / Total Sum of Squares)

Deployment

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., pick up point, time) and receive a cancellation prediction.

Preliminary Results



The classes are not balanced.

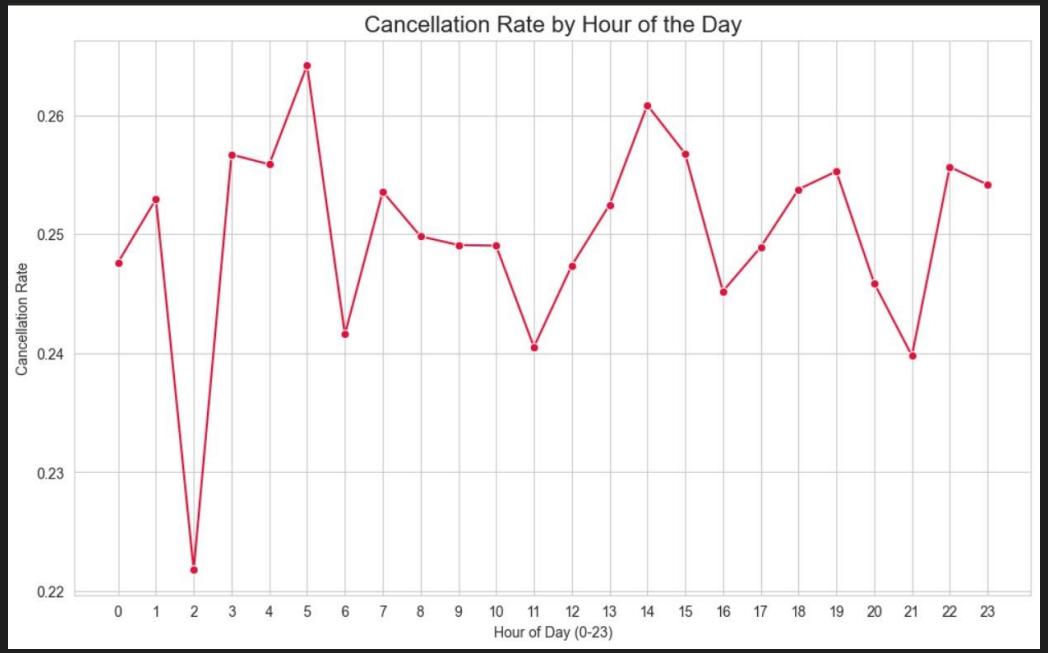
```
df.isna().sum()
✓ 0.0s
```

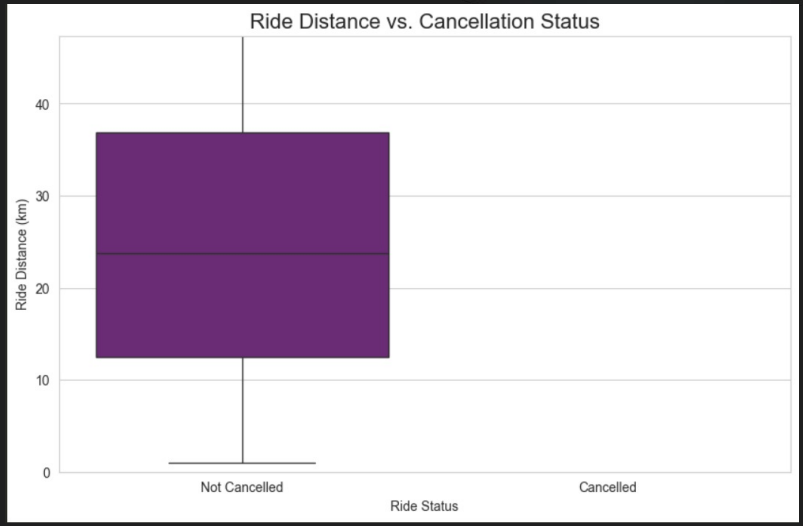
Date	0
Time	0
Booking ID	0
Booking Status	0
Customer ID	0
Vehicle Type	0
Pickup Location	0
Drop Location	0
Avg VTAT	10500
Avg CTAT	48000
Cancelled Rides by Customer	139500
Reason for cancelling by Customer	139500
Cancelled Rides by Driver	123000
Driver Cancellation Reason	123000
Incomplete Rides	141000
Incomplete Rides Reason	141000
Booking Value	48000
Ride Distance	48000
Driver Ratings	57000
Customer Rating	57000
Payment Method	48000

dtype: int64

NaN in one-hot-encoding.

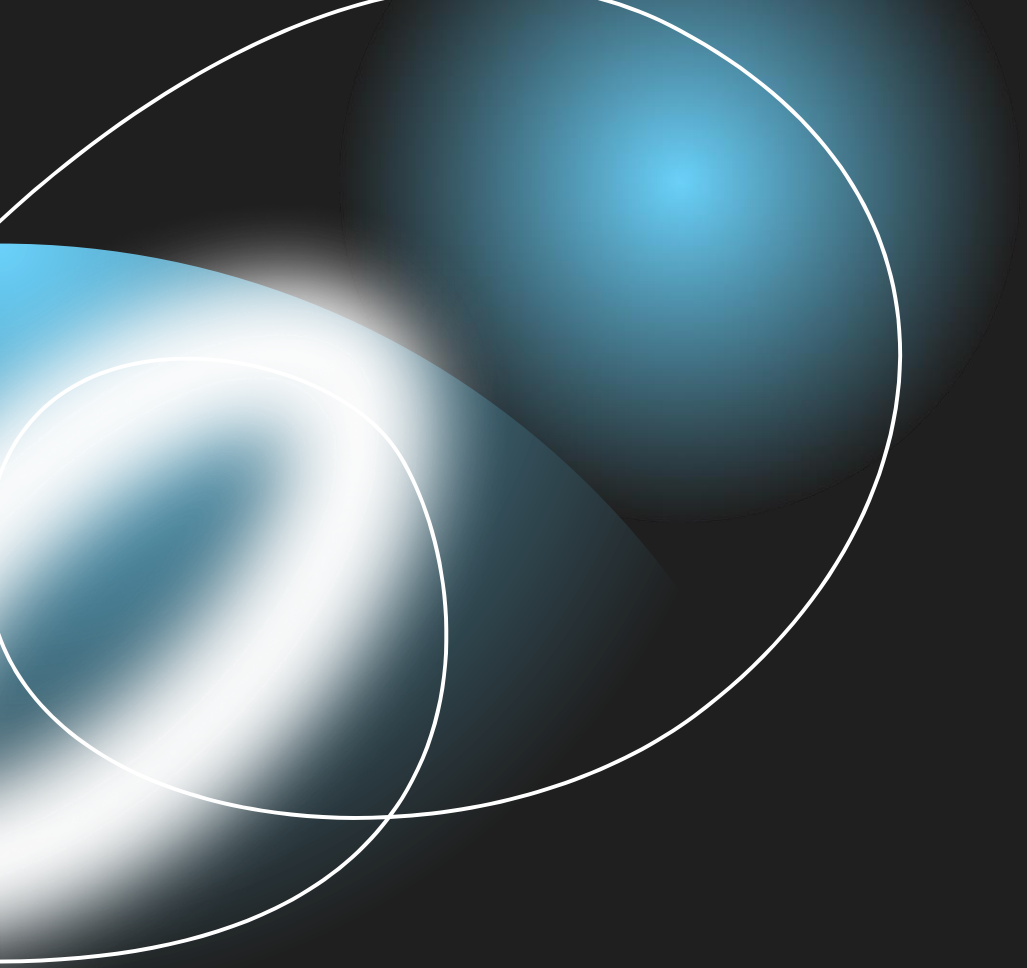
Preliminary Results





Cancellation rates seems to be random and irrelevant to time except when it goes down past midnight.

Missing distance between pickup and destination for cancelled rides.



Thank you!