# Machine Learning 2

# BIA-5402-0LA

Group Assignment 1

Taxi Cancellations Predictive Modeling Using Neural Networks

**Group:** 2

**Team Members:**

1. Alexandra Gladkova
2. Christina Saju
3. Hao Lun Rong
4. Kushwanth Sai Kolli
5. Ladan Asempour

**Submission Date**: 2/2/2025

1. **Introduction**

This report contains an analysis on a taxi booking dataset containing over 10,000 records from yourcabs.com, the report will go over a brief summary of the exploratory data analysis and data cleaning, and then focus on the key findings of the project based on several machine learning models.

1. **Exploratory Data Analysis**

**Cancellation Rate**

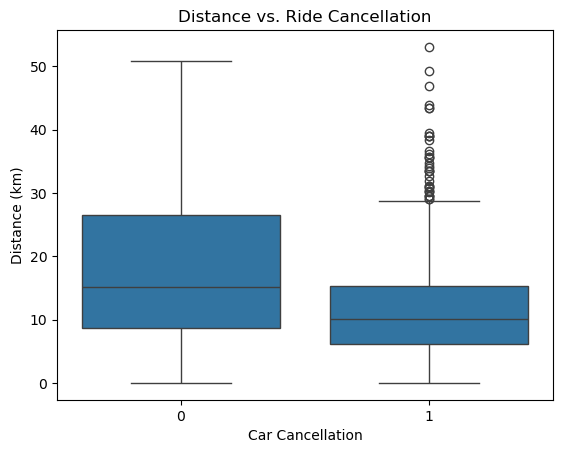
Before starting analysis, the cancellation rate was visualized. The dataset contained a total of 10,000 records, with which there were 743 cancelled rides.



1. **Data Cleaning & Feature Engineering**

Data cleaning was performed on the dataset by assessing what columns contained missing values as well as removing columns such as row# which were irrelevant to the analysis. We noticed large portions of data missing in certain columns such as package\_id, from\_cityid, to\_city id most notably. With over 70% of data missing for those columns we decided to remove them completely. The to\_date column had over 40% so it was removed, the from\_date column as taxi rides are likely the same day and would be useful for feature engineering, where the month, and day were extracted into separate features, the year column was not used as all data was in 2013.

The latitude and longitude data were then cleaned by removing null values for calculation. The geodesic distance was calculated from the latitudes and longitudes provided in the dataset. Plotting the relation between distance and ride cancellation, we found cancellations more likely when travel distance was further, however we decided to not use the distance calculation in our models as it would not reflect actual road routes.

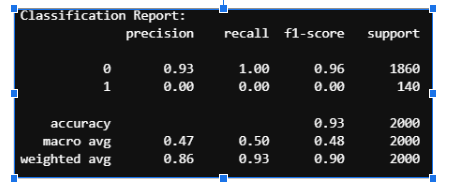
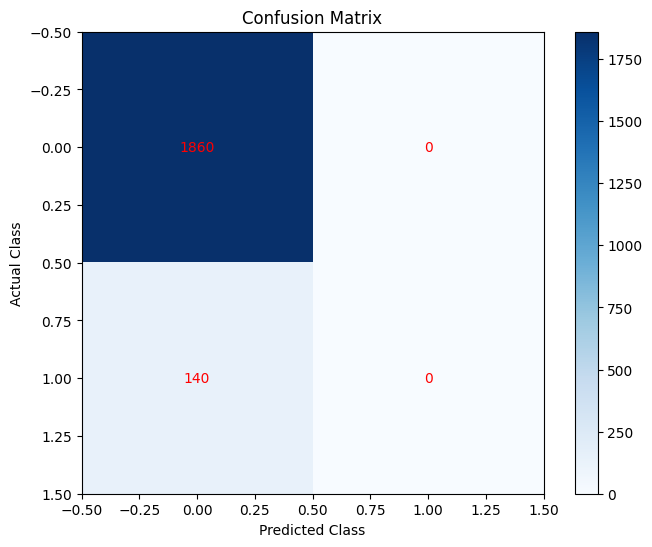


For the remaining features, vehicle\_model\_id and travel\_type\_id were encoded into dummy variables for each type.

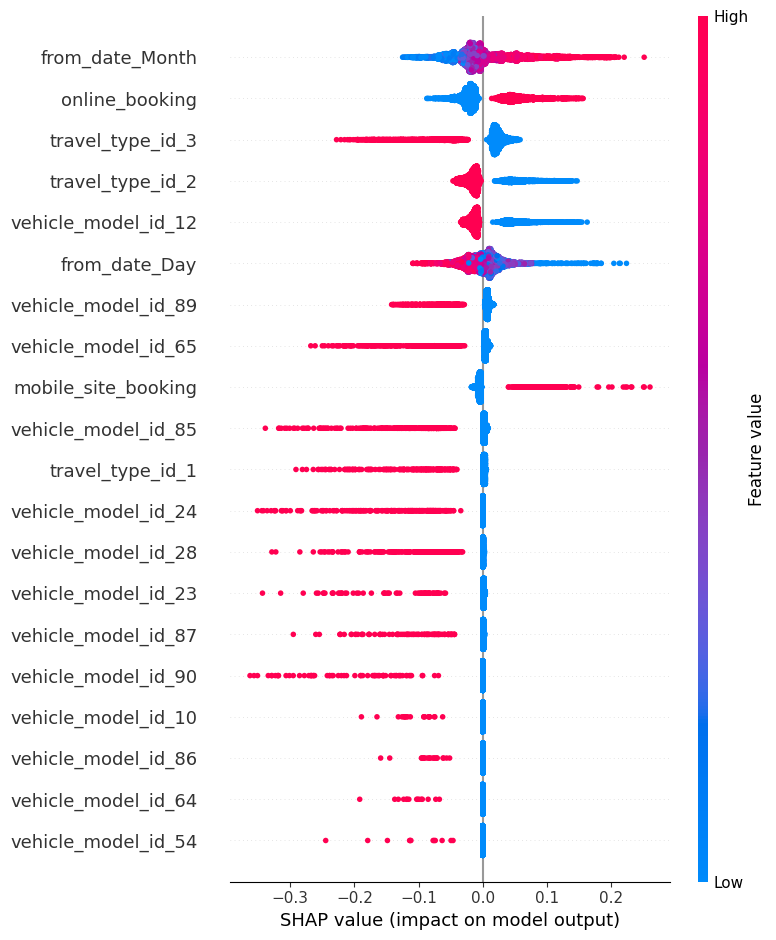
1. **Key Findings**

Running our neural network model with a run of 20 epochs, we have a test accuracy of 93%. However with the Confusion Matrix and Classification report we can see some very misleading results.

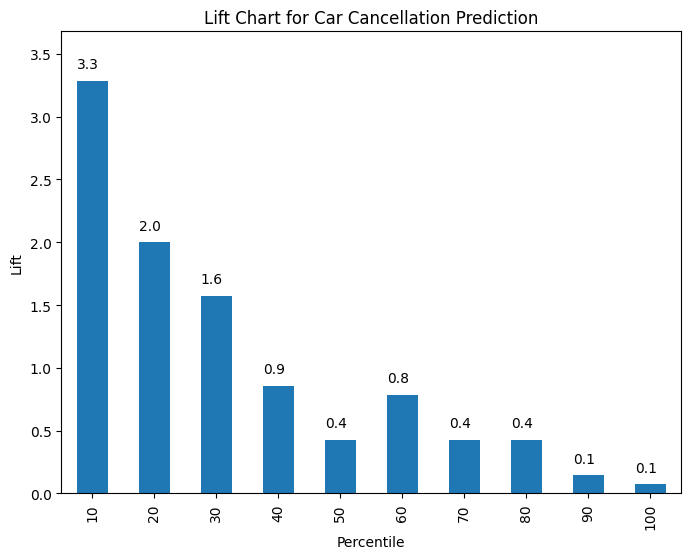
The model was able to accurately predict all classifications of non cancellations, however was not able to classify any cancellations correctly, this can be seen with great precision, recall, and f1-scores for the non cancellation classification, but a 0 score on anything cancelled. Further discussion of the issue will be in the recommendations and conclusions.



We also plotted the impact of each feature on the model output. The SHAP value graph shows each feature and its data points and how it affects the model. We can see that there are a lot of features which negatively affect the data in the form of dummy variables.



Lastly, a lift chart was plotted to help evaluate how well the model ranks high risk cases compared to a random selection. We can see from the chart that the model does a very good job in the top 30% but drops significantly after that. Which means if this model is used for targeted intervention then we should focus on the top 30% only for proactive actions. Beyond that there is a lot of room for improvement of the model.



1. **Improvements to the model and recommendations**

From our model’s performance metrics, our recommendations would be to remove the dummy variables which were created in vehicle\_type and travel\_type as we can tell the dummy variables are negatively impacting the model. We need more instances of rides cancelled for better insights. In addition, while the model ran to the constraints given in the project outline, we could increase the number of epochs to get better results, as well as changing some parameters in the neural network model.

As for recommendations to the company to reduce cancellations. When visualizing the relation between the columns and cancellations, we noticed that package\_id and vehiclemodel\_id had some effect on the number of correlations. As we do not know exactly what the identifications for these mean, we are unable to say for certain but it is speculated that users may be more likely to cancel if the order was based on a different type of travel such as carpool as opposed to ordering your own ride. As well as different cancellations based on the possible size of the vehicle among other things.   
Our recommendations to help with this would be to streamline their services more to reduce the amount of vehicle types there are, in addition to possibly reducing the cancellation time allowed based on the type of service in order to retain ride orders.