NYC Uber Tips Analysis

City Driving
Jintong Chang
Yufan Ding
Yijia Gong
Yitong Ma
Yifan Shen

Introduction

In the bustling city of New York, once dominated by yellow cabs, Uber has become a transformative player in transportation. This analysis aims to enhance our understanding of tipping dynamics within the realm of Uber services in New York City. Beyond being a simple gratuity, each tip serves as a narrative, offering insights into driver-passenger interactions. Our analysis involved exploratory data analysis, data visualization, and model analysis for tipping behavior prediction. The experiment shows that our final model selection can provide valuable insights into the factors that make customers give tips, which provide support for the business recommendations. Last but not least, we also discuss some limitations and further improvements to our model.

Data Cleaning and Processing

Data Introduction Our dataset comprises three sets of data. The largest set pertains to the high-volume FHV trip information for the 12 months of 2021. Each row represents an individual trip in an FHV dispatched by one of NYC's licensed high-volume FHV bases, and all files are in .parquet format. Initially, we converted the files for the 12 months into pandas data frames using the 'pyarrow' package. Subsequently, we filtered trips only with license number "HV0003," corresponding to Uber trip data. The variables used in our analysis include pickup_datetime, dropoff_datetime, trip_times, trip_miles, base_passenger_fare, sales_tax, tips, PULocationID, and DOlocationID. The second data file contains daily weather information for the year 2021, and our analysis focused on the variables temp, precip, and snow. The final data file concerns zone data, with LocationID and Borough being the key variables used in our analysis. These variables were employed to match the pick-up and drop-off IDs in the trip dataset. (Appendix: datasets)

Data Preprocessing Initially, we attempted to consolidate all 12 months of Uber trip data into a comprehensive and sizable data frame for analysis. However, the combined data frame included up to 100 million individual trips. Recognizing the need for efficiency in analysis, we decided to extract a representative subset of the data through random sampling, limiting it to 2000 trips. This approach allows us to conduct a thorough analysis while retaining the statistical characteristics of Uber trip data. Moreover, a significant portion of the dataset demonstrates sparse tipping behavior, with approximately 14%-15% of instances featuring tips. Consequently, we converted the 'tips' column into a categorical type. We believe that this transformation will enhance the interpretability of tipping patterns (Appendix: Figure1).

EDA Our EDA section involved conducting basic descriptive statistics and visualizations to understand customer behavior when using Uber. In terms of descriptive analysis, the mean trip time is approximately 18 minutes, and the average trip covers about 4 miles. This implies that people often opt for Uber for quick travel needs. We also examined the correlation between independent variables and observed a strong positive relationship between trip miles, trip times, and base passenger fare (Appendix: Figure 3). Despite this correlation, we opted to retain all of these variables as independent variables. This decision allows for a more comprehensive understanding of the factors influencing tipping behavior, enabling a thorough and nuanced analysis. To gain a deeper insight into customer behavior, we created visualizations for our dataset. Firstly, we explored the distribution of pickup and dropoff boroughs (Appendix: Figure 4). Manhattan emerged as the most common location for both pickup and dropoff, followed by Brooklyn. To enhance comprehension, we illustrated the distribution of pickup and dropoff hours throughout the day (Appendix: Figure 5). The majority of trips occurred in the afternoon and evening hours, providing a clearer picture of the temporal patterns in customer behavior.

Model Preparation

Model Specification In the course of implementing machine learning models for the prediction of tip probability, a comprehensive approach was undertaken, encompassing feature selection, feature engineering, and addressing class imbalance. The dataset initially comprised 11 independent variables, from which two variables, namely "dropoff datetime" and "sales tax," were removed to avoid multicollinearity because of the dataset's existing "pickup datetime" and "trip time" variables. Furthermore, to circumvent multicollinearity between "trip miles" and "trip time," an interaction term was introduced, denoted as the product of these two variables. In addition, measures were taken to prevent the dummy variable trap, notably by eliminating the borough of Staten Island from both pickup and drop-off locations. The purpose behind this elimination was to enhance the robustness and interpretability of the subsequent models.

Undersampling A challenge in the dataset is the huge imbalance in the classes, where the number of observations who tip is way less than the number of observations who did not tip. To solve this imbalance, an undersampling method was applied to the training dataset. This strategic undersampling aimed to equalize the representation of instances belonging to both tipping and non-tipping categories during the model training phase. We will compare the results from both cases in the model analysis. We refer to samples without undersampling as uncleaned data, and samples with undersampling as cleaned data.

Model Analysis

Logistic Regression The Logistic Regression model utilizes a linear combination of input features to estimate the probability of giving tips. After applying the model, we managed to construct the roc curve which gives us the best probability threshold for classification. We then applied this threshold in the test set predictions and constructed a confusion matrix. For our analysis: uncleaned data with the best threshold of 0.196, accuracy of 0.707, precision of 0.187, recall of 0.292, and f-score of 0.228 (Appendix: Figures 6&7); cleaned data with the best threshold of 0.560, accuracy of 0.803, precision of 0.216, recall of 0.124, f-score of 0.157 (Appendix: Figures 8&9).

Random Forest Random Forest, an ensemble learning method, aggregates predictions from multiple decision trees to enhance accuracy and robustness. Notably, Random Forest provides insights into feature importance, highlighting the variables that significantly contribute to the model's predictive performance. For our analysis, the top 3 variables are “dropoff_datetime”, “DOBorough_3" and “trip_time” (Appendix: Figure10). For our analysis: uncleaned data with the best threshold of 0.19, accuracy of 0.612, precision of 0.179, recall of 0.449, and f-score of 0.256 (Appendix: Figure 11&12); cleaned data with the best threshold of 0.52, accuracy of 0.543, precision of 0.161, recall of 0.494, and f-score of 0.243 (Appendix: Figure 13&14).

KNN K-Nearest Neighbors classifies data points based on the majority class of their K-nearest neighbors. In the KNN model, we first determined the value of neighbors by comparing the out-of-sample r-squared. Then we applied the model to the test set and got the results. For our analysis: uncleaned data with the best value of neighbors k=7, accuracy of 0.843, precision of 0, recall of 0, and f-score of 0 (Appendix: Figure 15&16); cleaned data with the best value of neighbors k=12, accuracy of 0.525, precision of 0.120, recall of 0.348, and f-score of 0.179 (Appendix: Figure 17&18).

Decision Tree The decision tree model makes predictions by recursively splitting the data based on features, forming a tree structure. We once again used the resampled training data to tune the model. More specifically, we employed 10-fold cross-validation and grid search to identify the best hyperparameters, such as the maximum depth of the tree and the criterion for splitting. Eventually, we retrieved the results by implementing our model on the test set. For our analysis: uncleaned data with an accuracy of 0.852, precision of 0, recall of 0, and f-score of 0 (Appendix: Figure19); cleaned data with an accuracy of 0.480, precision of 0.155, recall of 0.562, and f-score of 0.243 (Appendix: Figure20).

Model Comparison and Selection In evaluating the performance of different models on the undersampling data, we emphasized both accuracy and precision as crucial metrics in selecting the ideal model for predicting tips. While accuracy measures overall correctness, precision specifically gauges the accuracy of positive predictions, for example, to ensure that drivers are not erroneously expecting tips. Given the significance of correctly identifying instances of tips, a model striking a balance between accuracy and precision proves to be the most favorable.

Based on the results, the logistic regression has the best accuracy and precision among all the models to predict whether the passenger is going to tip. At the same time, the model has high explainability in which we can access the coefficient of each independent variable and figure out their impact on the probability. Random forests also have the second-best accuracy and precision (Appendix: Figure 21&22). Even though the explainability of random forest is less than the logistic regression. We were able to draw out the features’ importance from the model (Appendix: Figure10). Additionally, the uncleaned data exhibited a higher accuracy and some invalid zeros compared to the cleaned data, suggesting that undersampling can yield better and less biased results. (Appendix: Figure 23)

Business Recommendation

Based on our model analysis and insights into tipping behavior within the Uber community, we recommend strategic interventions to enhance tipping dynamics. Our first proposal involves incentivizing tipping awareness through targeted campaigns and in-app notifications, highlighting the positive impact of tips on driver livelihoods. Additionally, we advocate for educational initiatives to inform users about tipping factors, fostering transparency in how tips contribute to service quality. Introducing a reward program with tangible benefits, such as discounts or loyalty points, can reinforce positive tipping behavior and promote customer loyalty. Enhanced feedback mechanisms and surge tipping options during peak demand can align tipping behavior with service demands. Finally, a driver recognition program based on positive feedback and tipping patterns can motivate exceptional service, emphasizing the importance of tipping.

Connecting these recommendations to our model analysis, our findings indicate significant correlations between variables like trip duration, distance, base fare, and temperature with tipping behavior. This insight allows us to propose targeted actions, such as implementing automatic surge tipping reminders during rainy rush hours, leveraging our model-driven understanding to enhance user experiences and encourage positive tipping outcomes.

Limitations and Future Improvements

Firstly, the impact of limited or biased data on model performance is a primary concern. Incomplete or inaccurate data might lead to biased predictions. Notably, our model lacks crucial user and driver information like ratings and demographics, which could significantly influence tipping. To enhance predictions, incorporating relevant features such as customer behaviors and contact details is essential. Secondly, a tradeoff exists between model accuracy and interpretability, especially in complex models like neural networks and random forests. While high-accuracy models might suit certain predictive tasks, understanding feature influences on tipping behavior demands models with higher interpretability. Different models serve different purposes based on this tradeoff.

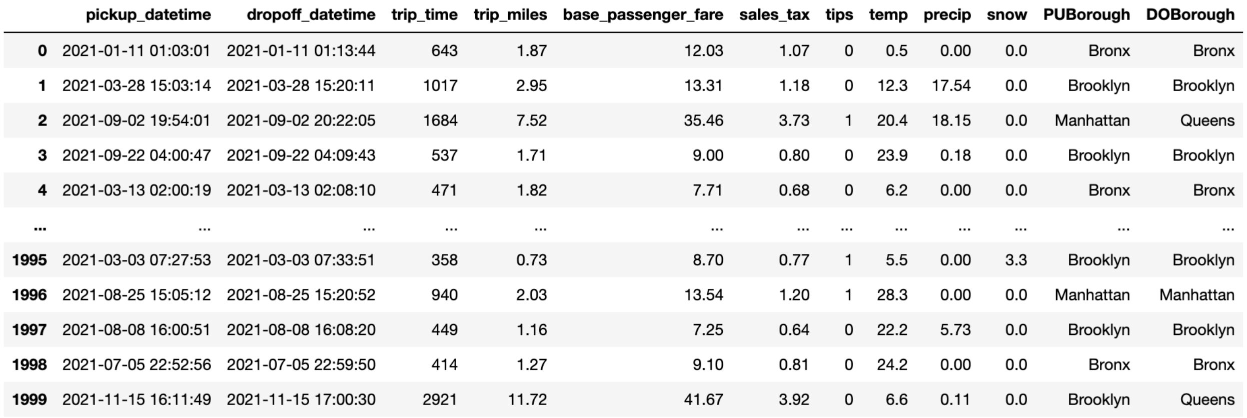
Meanwhile, the relevance of personalizing models based on individual preferences is evident. For instance, assigning more weight to weather-related features for customers inclined to tip in adverse weather could enhance accuracy. Leveraging text analysis from customer reviews offers insights into these preferences. Last but not least, considering the dataset's 2021 origin, the dynamic nature of industry trends and policies is crucial. Constantly updating models with sequential training and testing is essential to ensure relevance and accuracy amidst evolving industry dynamics.

Appendix

Datasets (original):

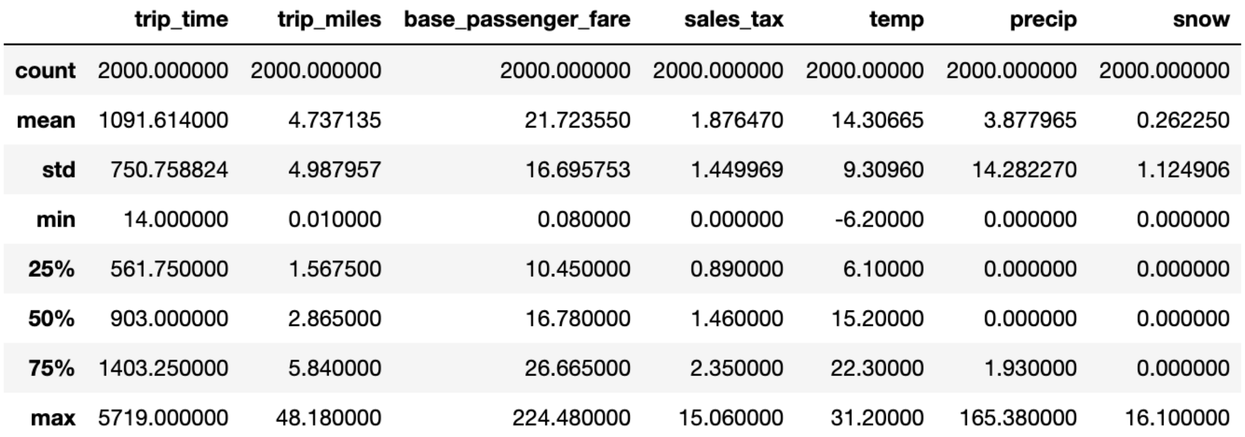
- shuheng_mo. (2023, February 2). *Uber NYC for-hire vehicles trip data (2021)*. Kaggle. https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021

Figures:



	pickup_datetime	dropoff_datetime	trip_time	trip_miles	base_passenger_fare	sales_tax	tips	temp	precip	snow	PUBorough	DOBorough
0	2021-01-11 01:03:01	2021-01-11 01:13:44	643	1.87	12.03	1.07	0	0.5	0.00	0.0	Bronx	Bronx
1	2021-03-28 15:03:14	2021-03-28 15:20:11	1017	2.95	13.31	1.18	0	12.3	17.54	0.0	Brooklyn	Brooklyn
2	2021-09-02 19:54:01	2021-09-02 20:22:05	1684	7.52	35.46	3.73	1	20.4	18.15	0.0	Manhattan	Queens
3	2021-09-22 04:00:47	2021-09-22 04:09:43	537	1.71	9.00	0.80	0	23.9	0.18	0.0	Brooklyn	Brooklyn
4	2021-03-13 02:00:19	2021-03-13 02:08:10	471	1.82	7.71	0.68	0	6.2	0.00	0.0	Bronx	Bronx
...	...	...	...	...	...	...	...	...	...	...	...	...
1995	2021-03-03 07:27:53	2021-03-03 07:33:51	358	0.73	8.70	0.77	1	5.5	0.00	3.3	Brooklyn	Brooklyn
1996	2021-08-25 15:05:12	2021-08-25 15:20:52	940	2.03	13.54	1.20	1	28.3	0.00	0.0	Manhattan	Manhattan
1997	2021-08-08 16:00:51	2021-08-08 16:08:20	449	1.16	7.25	0.64	0	22.2	5.73	0.0	Brooklyn	Brooklyn
1998	2021-07-05 22:52:56	2021-07-05 22:59:50	414	1.27	9.10	0.81	0	24.2	0.00	0.0	Bronx	Bronx
1999	2021-11-15 16:11:49	2021-11-15 17:00:30	2921	11.72	41.67	3.92	0	6.6	0.11	0.0	Brooklyn	Queens

Figure 1: Final Dataframe



	trip_time	trip_miles	base_passenger_fare	sales_tax	temp	precip	snow
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.00000	2000.000000	2000.000000
mean	1091.614000	4.737135	21.723550	1.876470	14.30665	3.877965	0.262250
std	750.758824	4.987957	16.695753	1.449969	9.30960	14.282270	1.124906
min	14.000000	0.010000	0.080000	0.000000	-6.20000	0.000000	0.000000
25%	561.750000	1.567500	10.450000	0.890000	6.10000	0.000000	0.000000
50%	903.000000	2.865000	16.780000	1.460000	15.20000	0.000000	0.000000
75%	1403.250000	5.840000	26.665000	2.350000	22.30000	1.930000	0.000000
max	5719.000000	48.180000	224.480000	15.060000	31.20000	165.380000	16.100000

Figure 2: Descriptive Statistics

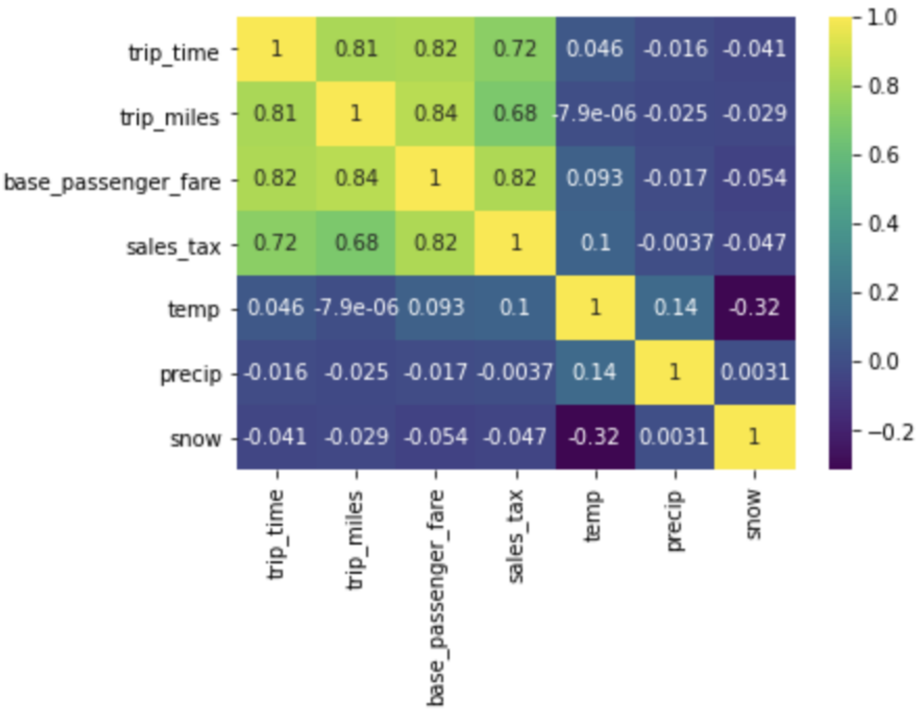


Figure 3: Correlation Graph

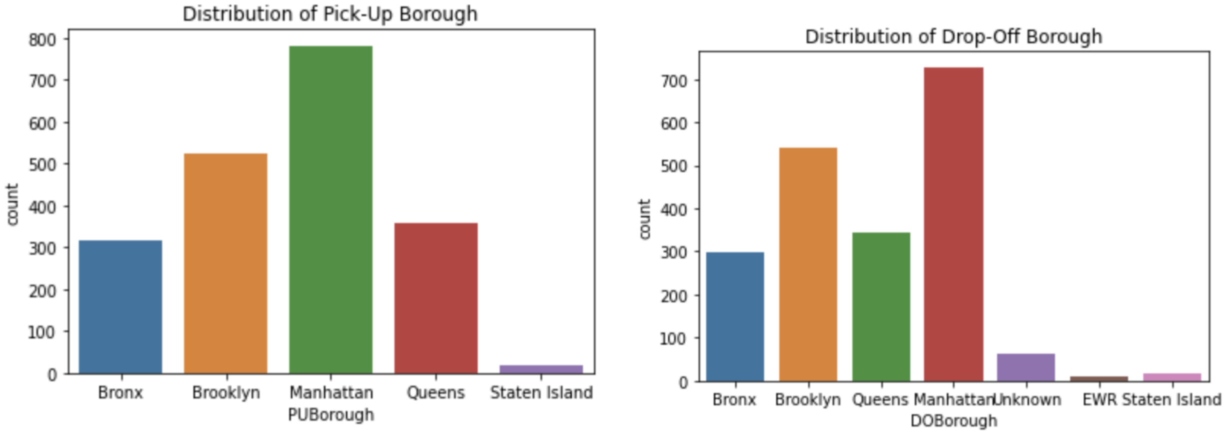


Figure 4: Distribution of Pick-Up and Drop-Off Borough

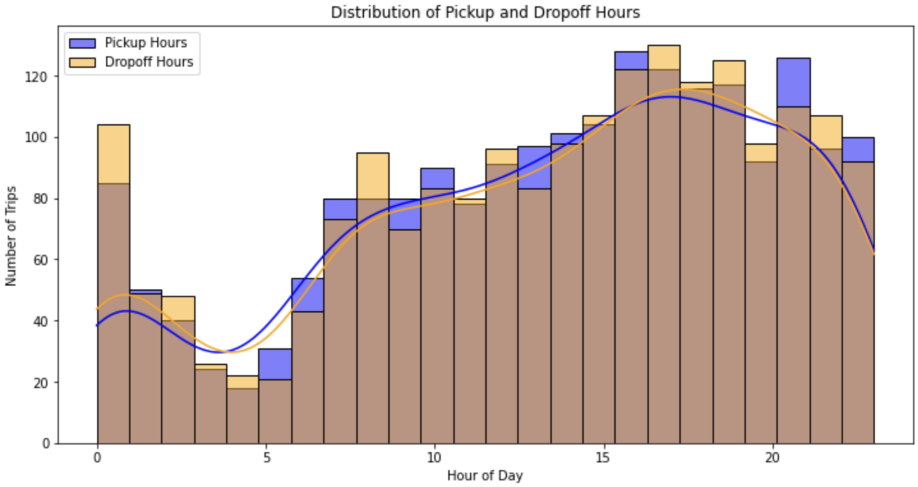


Figure 5: Distribution of Pick-Up and Drop-Off Hours

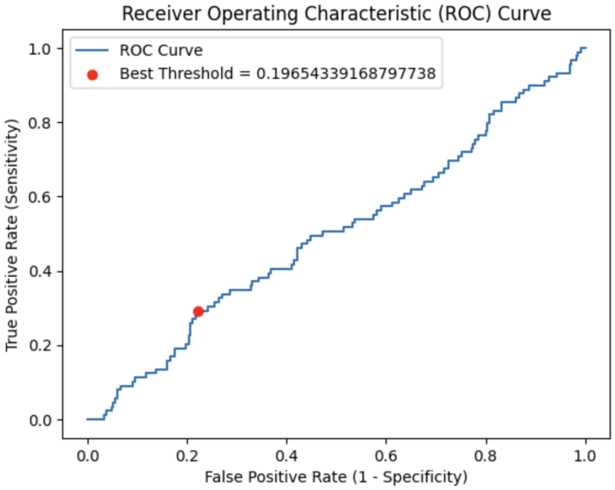


Figure 6: Logistic Regression (uncleaned) - ROC Curve

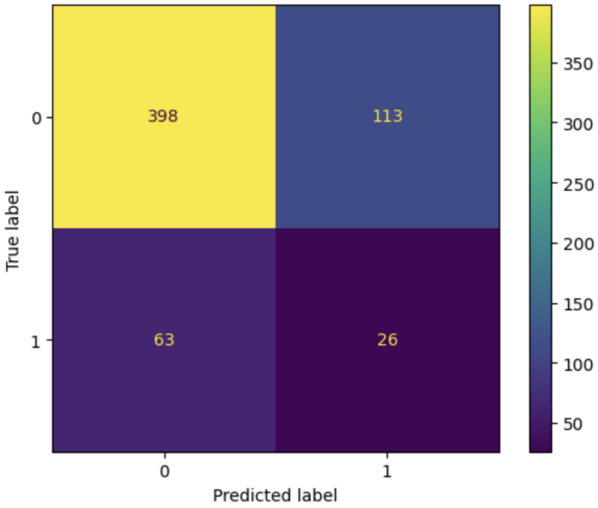


Figure 7: Logistic Regression (uncleaned) - Confusion Matrix

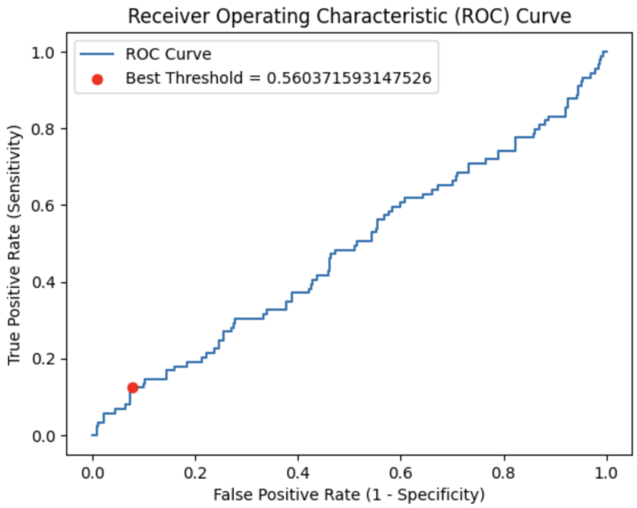


Figure 8: Logistic Regression (cleaned) - ROC Curve

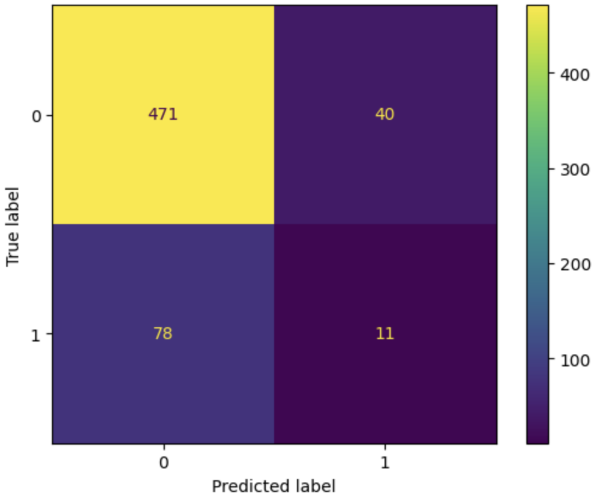


Figure 9: Logistic Regression (cleaned) - Confusion Matrix

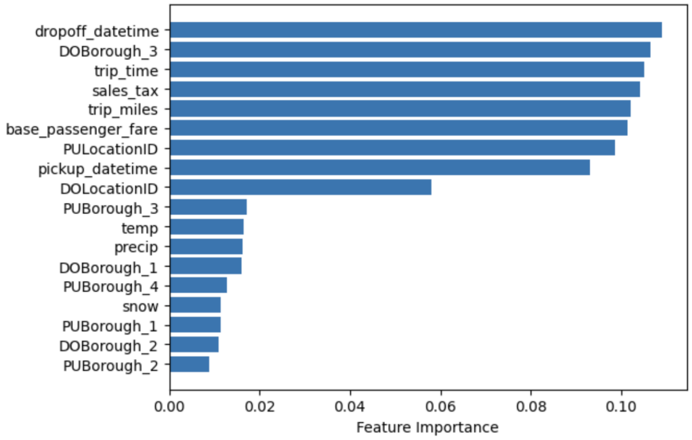


Figure 10: Random Forest (cleaned) - Feature Importance

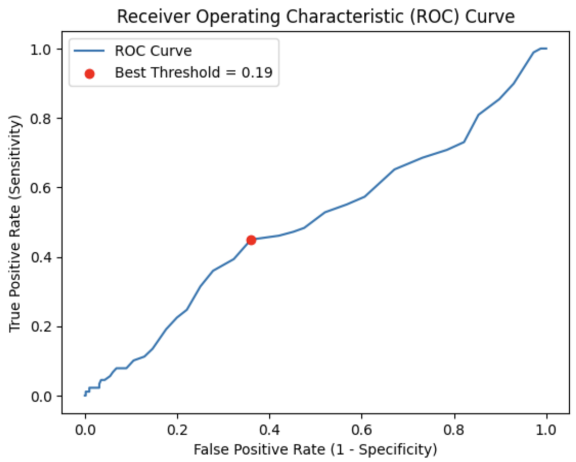


Figure 11: Random Forest (uncleaned) - ROC Curve

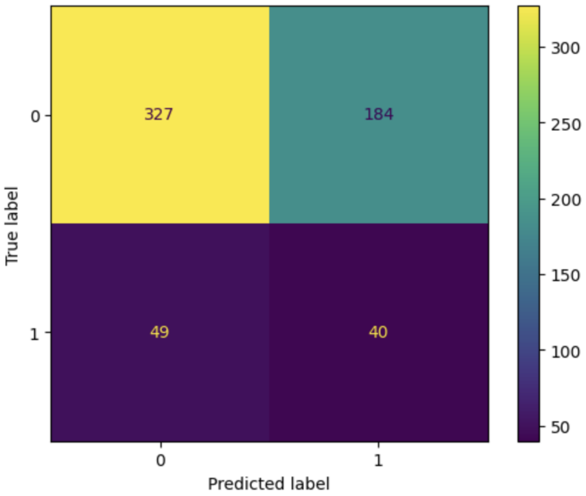


Figure 12: Random Forest (uncleared) - Confusion Matrix

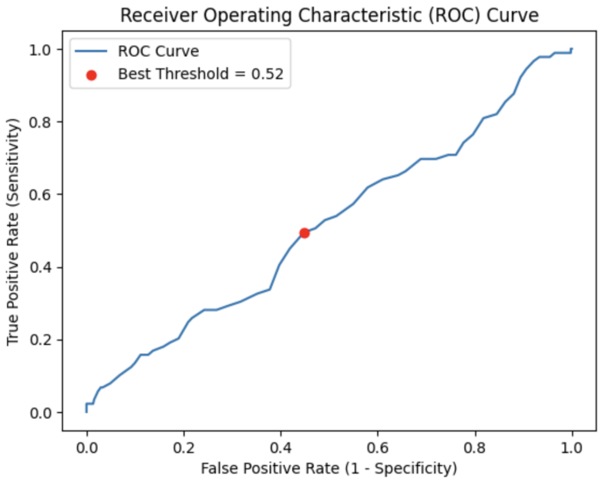


Figure 13: Random Forest (cleaned) - ROC Curve

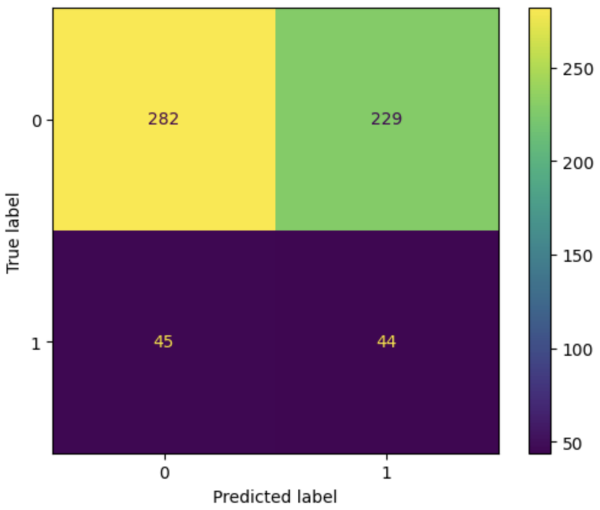


Figure 14: Random Forest (cleaned) - Confusion Matrix

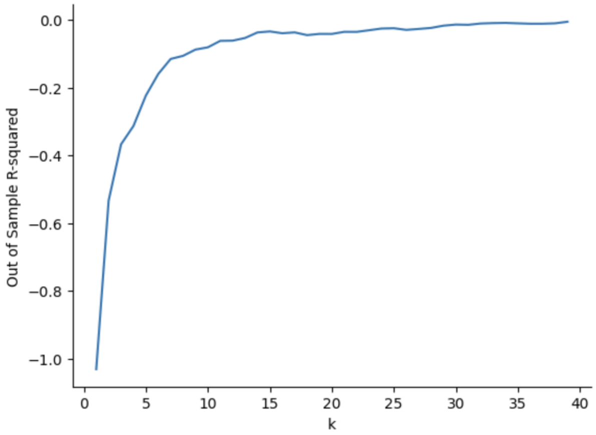


Figure 15: KNN (uncleaned) - K Neighbors

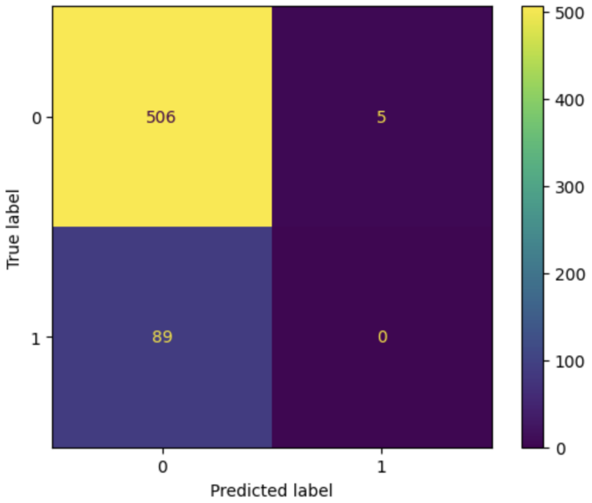


Figure 16: KNN (uncleaned) - Confusion Matrix

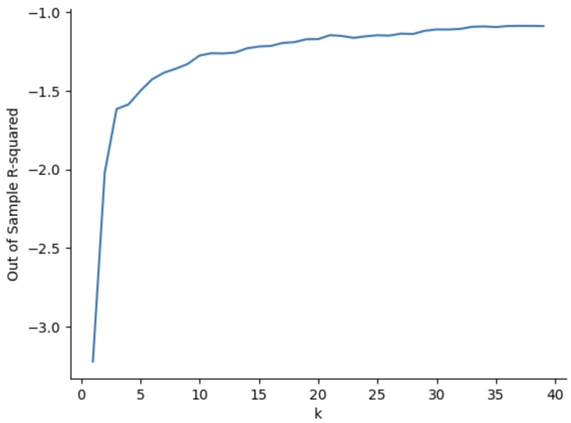


Figure 17: KNN (cleaned) - K Neighbors

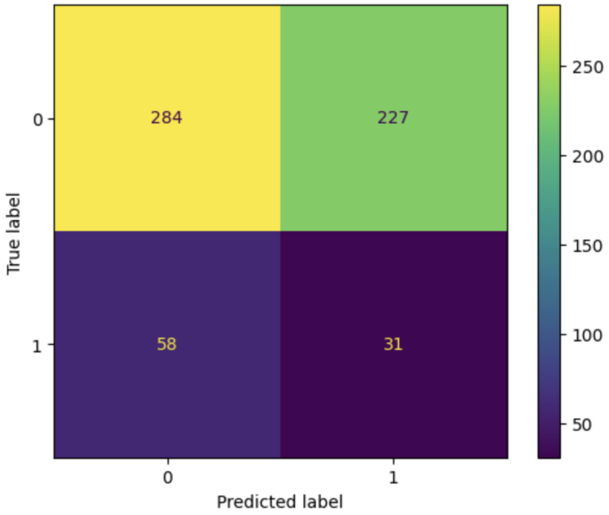


Figure 18: KNN (cleaned) - Confusion Matrix

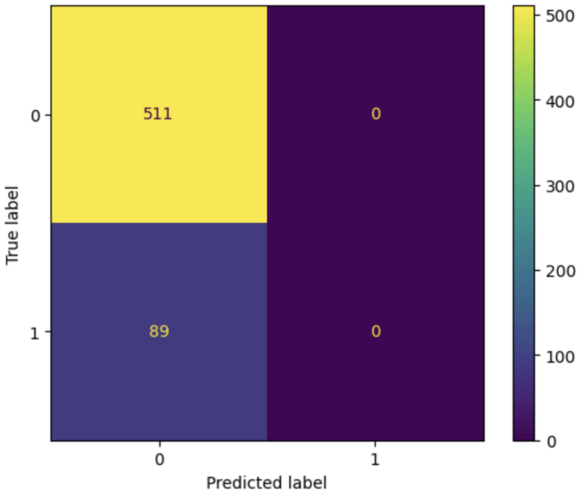


Figure 19: Decision Tree (uncleaned) - Confusion Matrix

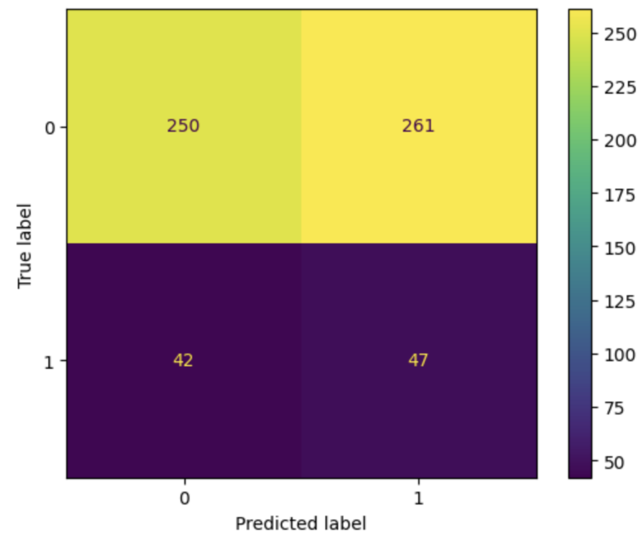


Figure 20: Decision Tree (cleaned) - Confusion Matrix

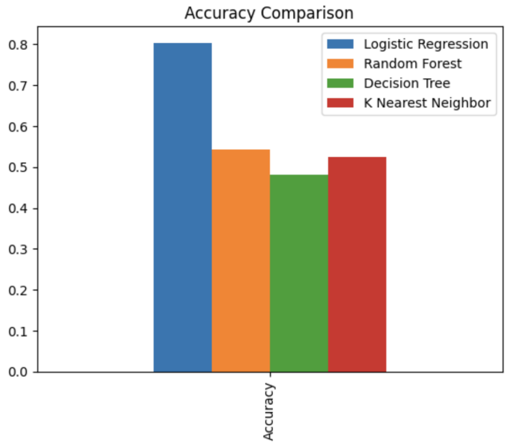
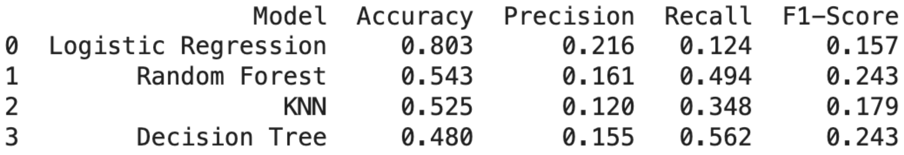
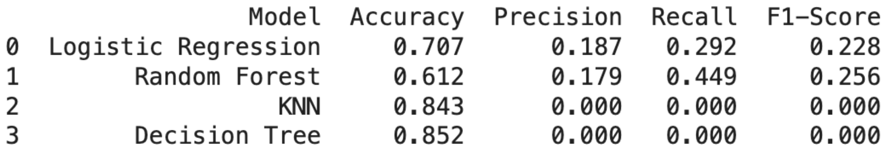


Figure 21: Accuracy Comparison



```
                 Model  Accuracy  Precision  Recall  F1-Score
0  Logistic Regression     0.803      0.216   0.124     0.157
1        Random Forest     0.543      0.161   0.494     0.243
2                  KNN     0.525      0.120   0.348     0.179
3        Decision Tree     0.480      0.155   0.562     0.243
```

Figure 22: Undersampling Model Score Results



```
                 Model  Accuracy  Precision  Recall  F1-Score
0  Logistic Regression     0.707      0.187   0.292     0.228
1        Random Forest     0.612      0.179   0.449     0.256
2                  KNN     0.843      0.000   0.000     0.000
3        Decision Tree     0.852      0.000   0.000     0.000
```

Figure 23: Uncleaned Model Score Results