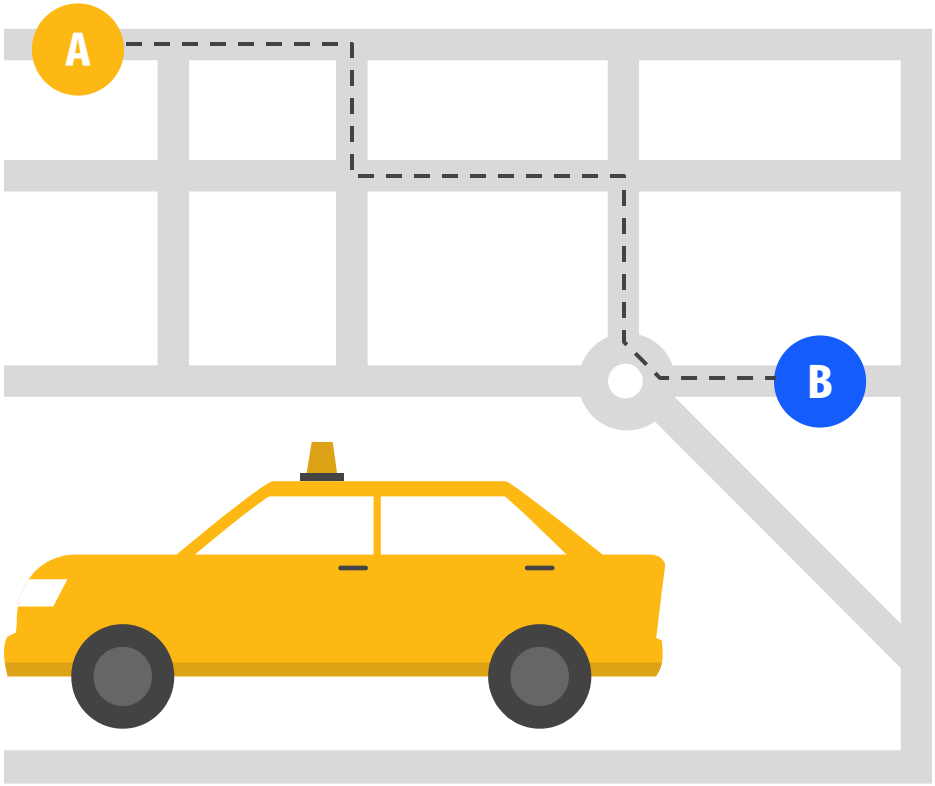
Where Do All the Cabs Go in the Late Afternoon?

Maira Asmat

From 4 p.m. to 5 p.m.

The number of active taxis falls by 20%

Problem



Model Purpose:
Interpret the importance of the features on the target to advise NYC on if this time should be changed

Classification Metrics:

Accuracy

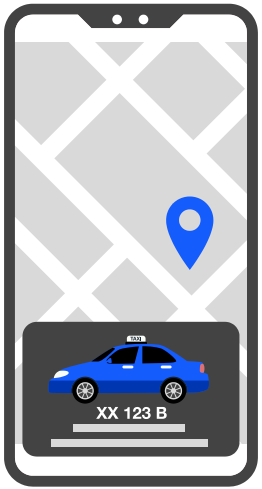
Dataset

1

N = 10,000

Obtained from official NYC taxi and limousine commission website

2



13 Features

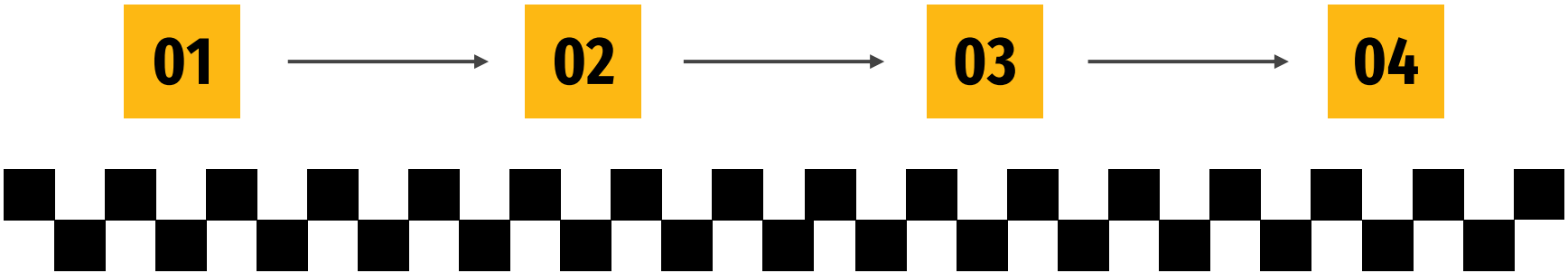
Passenger count, trip distance, location, various fees

3

Target

6 – 10 am (Morning) vs. 4 – 8 pm (Evening)

Methodology



Obtaining Data + EDA

Baselining and Choosing Models

Logistic Regression, Decision Tree, Random Forest

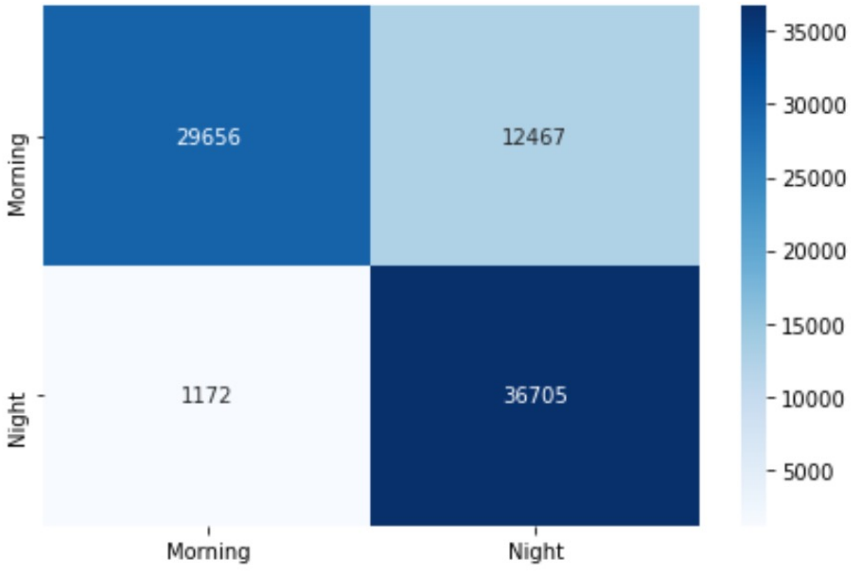
Tuning Model Hyperparameters

Using GridSearchCV on scaled data

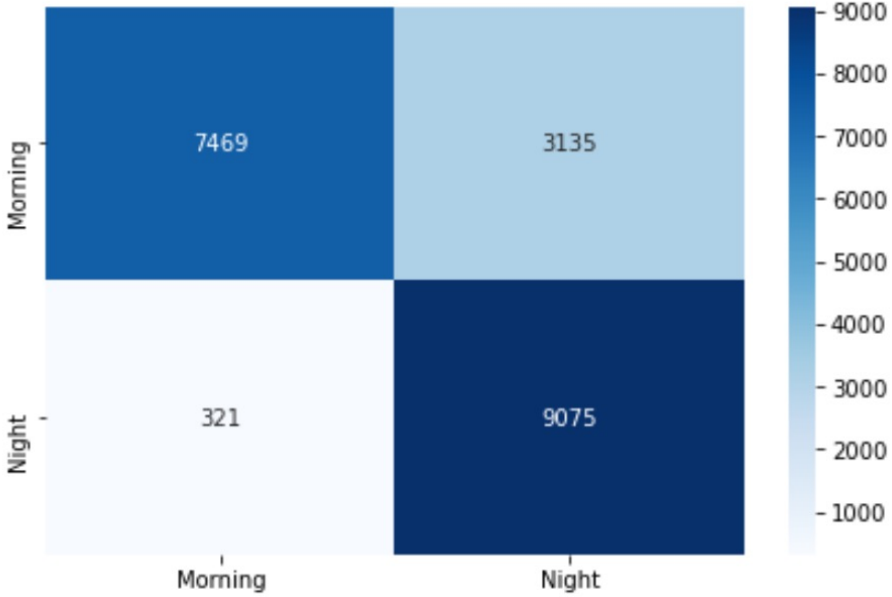
Interpreting Features

Final Model: Random Forest

- Train accuracy: 0.829



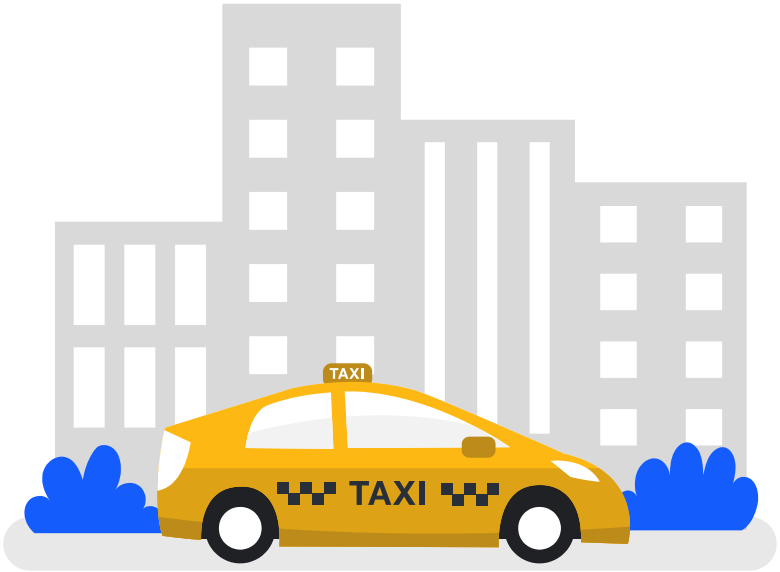
- Test Accuracy: 0.827



Interpreting Features

Most Important Features:

- Passenger Count
- Extra charges (Overnight and rush hour)
- Dropoff Location
- Total Amount



Least Important Features:

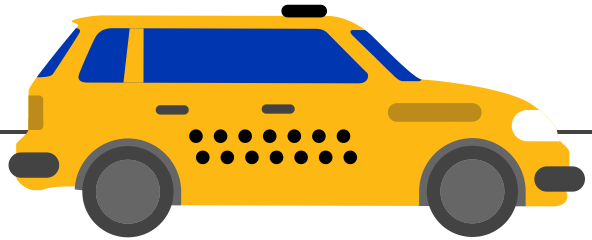
- Tolls
- MTA tax (based on metered rate in use)

Future Work

01

Do look into changing the shift time

02



Research more on how each feature changes between classes

03

Find a better method to determine tips

Appendix

	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.8175	0.8267	0.8274
Precision	0.9806	0.9631	0.9645
Recall	0.7153	0.7446	0.7446