DATASCI 420: TAXI TIP CLASSIFICATION

MAPLE TAN

OVERVIEW

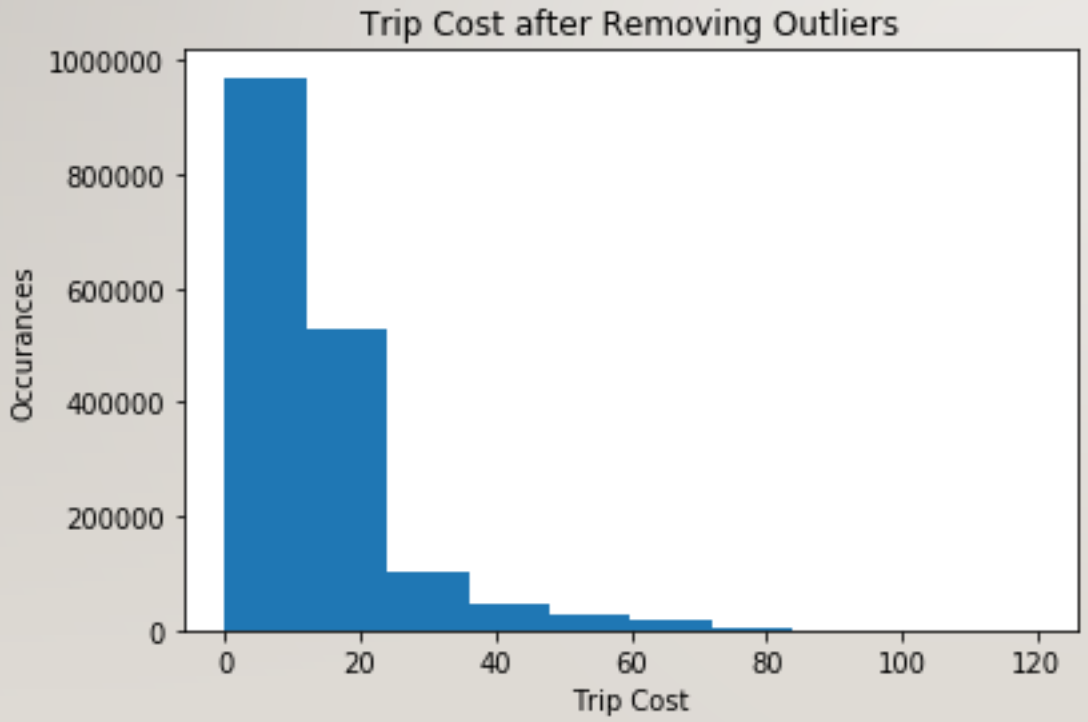
- Separated Taxi Tips into Low, Medium and High categories based on percentage of total cost
  - < 15% of Total Cost = Low Tip
  - >= 15% and >= 20% of Total Cost = Medium Tip
  - > 20% of Total Cost = High Tip
- Target Feature: Tip Category
- Goal was to create a classifier that could categorize each trip as one of the three

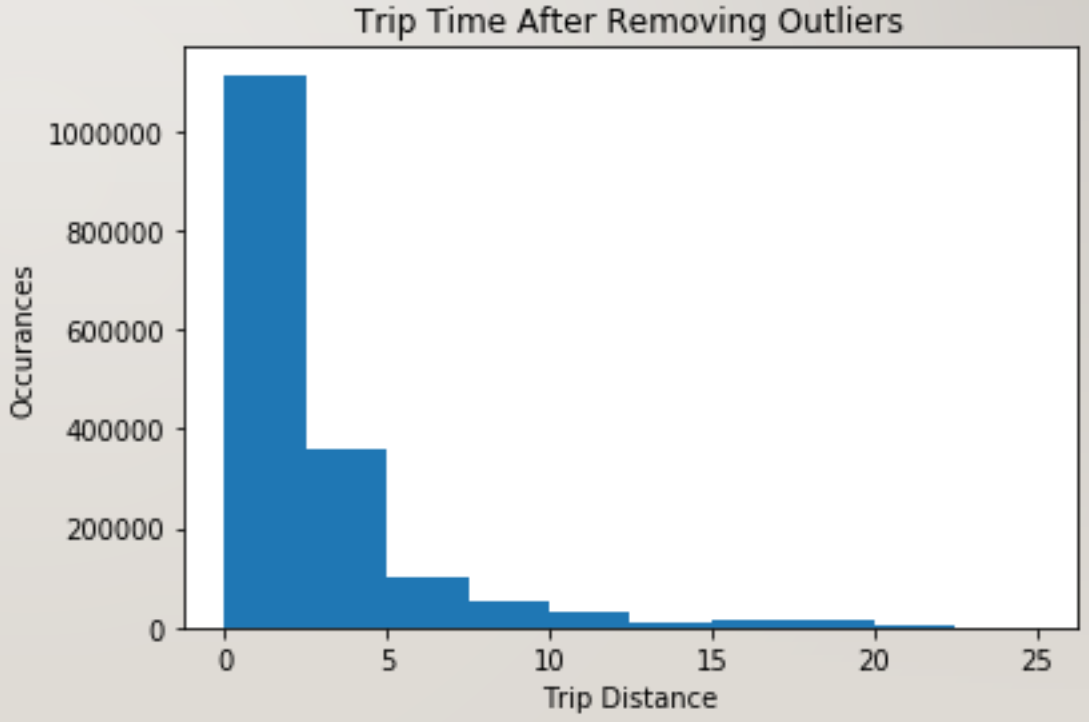
DATA

- 2013 New York Taxi Data from Class
  - Fare description
  - Trip Description
- 2013 New York Weather Data from National Centers for Environmental Information
  - Measured from various Weather Towers near the Central Park area
  - A decent chunk of missing data from first three months
- All data compiled into one giant data frame based on date and taxi trip start

DATA –CONTINUED

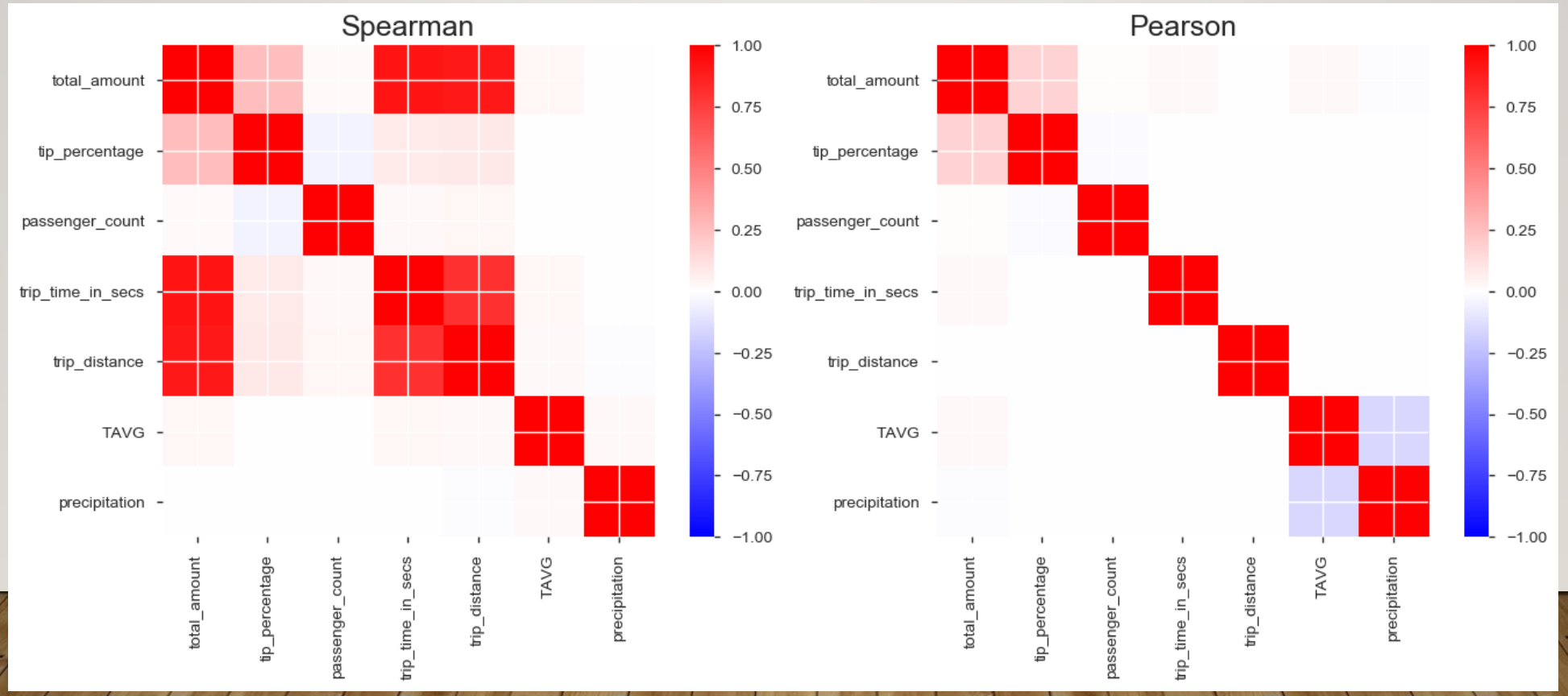
- Many features in Taxi Data have a long tailed distribution – basically all skewed right
  - Implies most trips were relatively short





DATA - CONTINUED

- Tip Percentage was slightly positively correlated with overall Trip Amount



FEATURE ENGINEERING

- Created a variable called ‘precipitation’ based off the rain and snow levels
  - Formula:
    - Precipitation = Rain + (Snow * 2)
- Created a column labelled ‘rush_hour’ which was True if trip started during rush hour (defined as 6AM – 10AM or 4PM – 8PM) or was False otherwise
- Used LabelEncoder on Categorical Variables to change them into numeric variables

FEATURES

Feature	Description
payment_type	Method customer used to pay for trip
total_amount	Total Amount paid on trip
passenger_count	Number of passengers on trip
trip_time_in_secs	Duration of trip in seconds
trip_distance	Distance of trip in miles
TAVG	Average temperature for the day
precipitation	Measure of how much snow and rain for the day
rush_hour	True if trip starts during rush hour, false otherwise

MODEL BUILDING

- Attempted to use 5 Different Models:
  - Naïve Bayes Classifier
  - Decision Tree Classifier
  - Random Forest Classifer
  - K Nearest Neighbors
  - SVM
- Naïve Bayes trained fastest but was least accurate
- SVM took over 30 minutes to train and so was eliminated

MODEL EVALUATION

- Metrics were fairly stable between different trials – only about a 2-3% fluctuation

Metrics of a Single Trial				
Model	Accuracy	Recall	Precision	Runtime
Naïve Bayes	0.389	0.335	0.296	A few seconds
DecisionTree	0.927	0.848	0.846	A minute or so
RandomForest	0.8	0.579	0.539	Several Minutes >5 mins
K Nearest Neighbor	0.774	0.563	0.559	A minute or so
SVM	n/a	n/a	n/a	Too Long to run >30 Mins

POSSIBLE IMPROVEMENTS

- Include Traffic Data such as congestion or
  - ‘rush_hour’ might not especially be applicable to New York
- Factor in pick up and drop off locations
- Try out Logistic Regression due to skewed nature of the data
- Check for Overfitting

THANK YOU!