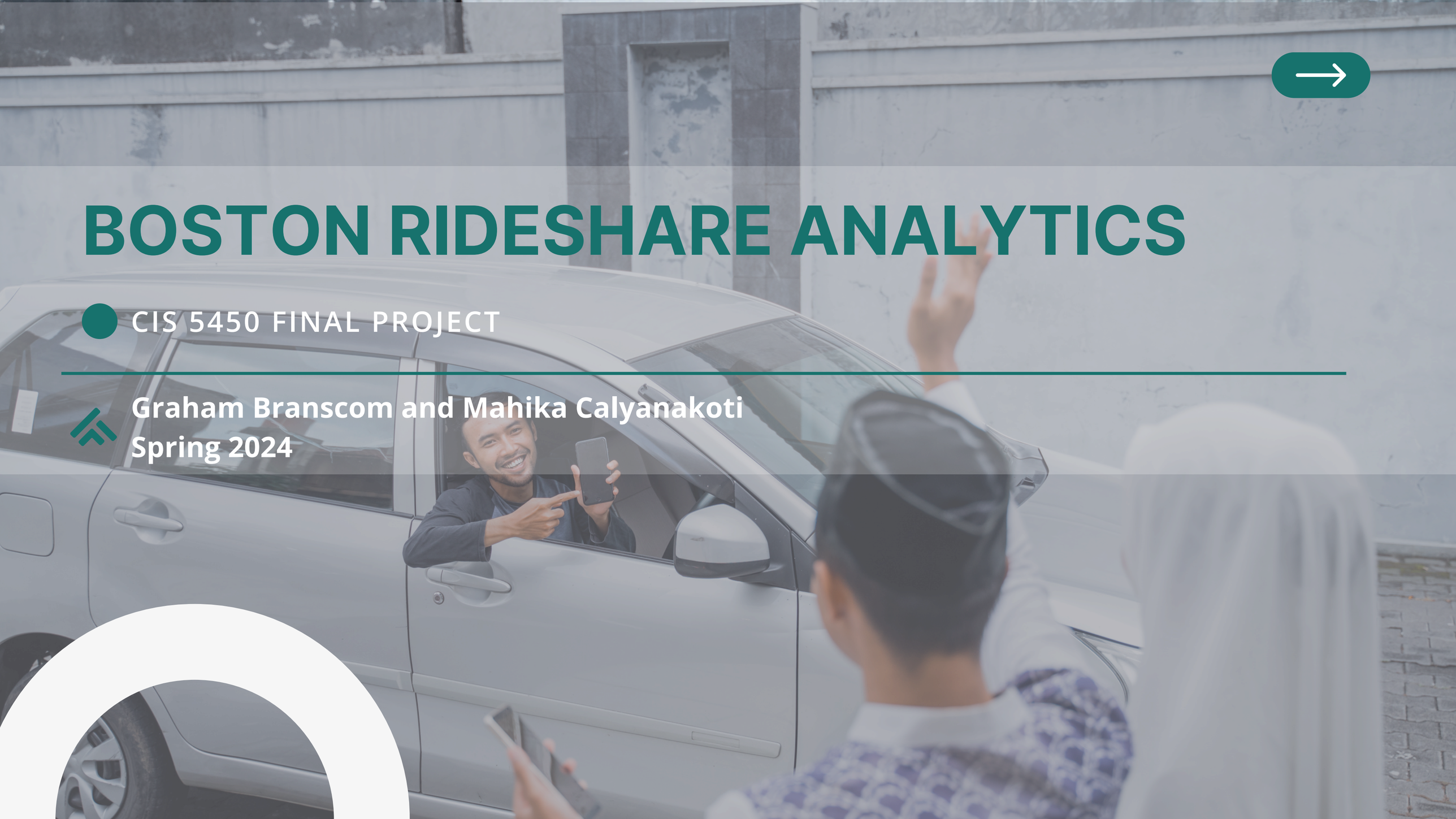


BOSTON RIDESHARE ANALYTICS

CIS 5450 FINAL PROJECT

Graham Branscom and Mahika Calyanakoti

Spring 2024

TABLE OF CONTENTS

FINAL PROJECT PRESENTATION

Overview

01

PROBLEM STATEMENT

INTRO, BACKGROUND, AND VALUE

With the rise of rideshare companies **revolutionizing transportation** in today's world, and with **college students** being a primary target customer base, we sought to analyze the various **factors impacting the costs** of these services by analyzing a **Kaggle** dataset on rideshare data from **Boston**, Massachusetts. With information on rides from Uber and Lyft, the two biggest rideshare companies in America, along with **weather data** based on the ride's location and time, this rich dataset allowed us to build complex models to gain **business and consumer insights** to see how the companies perform in comparison, and how consumers (especially students) can understand the pricing factors behind their favorite transportation services.

Goal:

1. Derive novel insights from this comprehensive dataset on **what factors**, both **direct** (such as rideshare company, distance, and time) and **indirect** (such as preciptation and temperature), **impact the price** of a given ride (regressional analysis).
2. Determine if we can construct a model that **classifies rides as either Uber or Lyft** (classification analysis)?.
3. Use our model results to paint a multi-dimensional picture of the ridesharing business landscape.

02

DATA & EDA

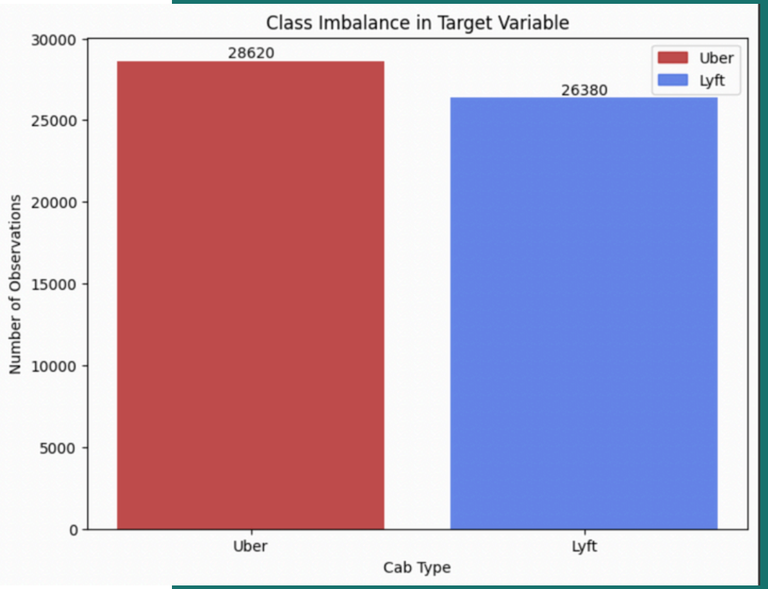
EXPLORING THE DATASET

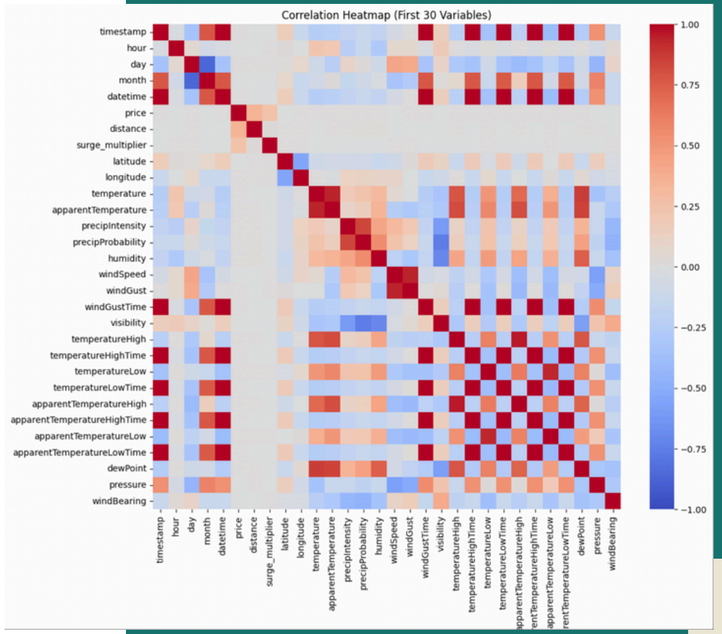
Dataset Description:

- 693,000 observations and 57 features
- Including the time, day, source, destination, distance, location, and rideshare company
- Weather features: temperature, short summary (e.g. overcast, rainy), precipitation, sunrise, and more
- Response Variable (Regression Models): Price in $USD
- Response Variable (Classification Models): Rideshare Company (Uber/Lyft)

Data Preparation and Feature Engineering:

- Dropped NA's, selected 55k subset, categorical type conversions, class **imbalance** in cab type (stratify)
- Created year and DayLength columns
- Applied **One-hot** encodings using **Regex** in preparation for correlation matrix
- Removed **highly correlated variables**, as well as outliers
- **Scale** data in prepration for future models



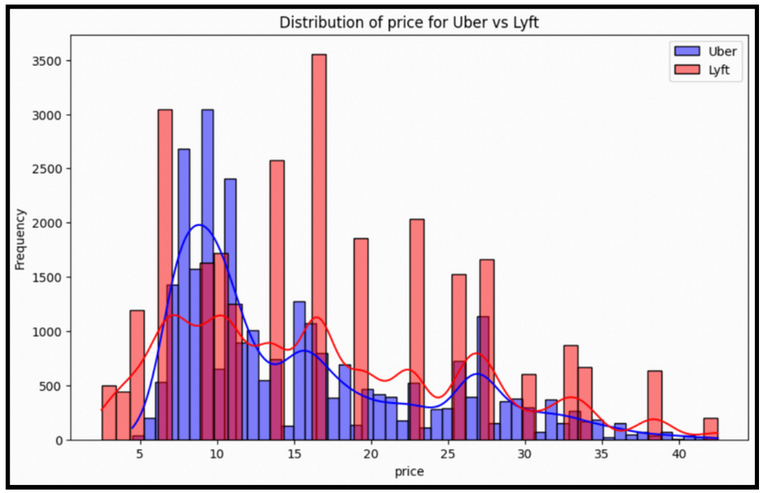


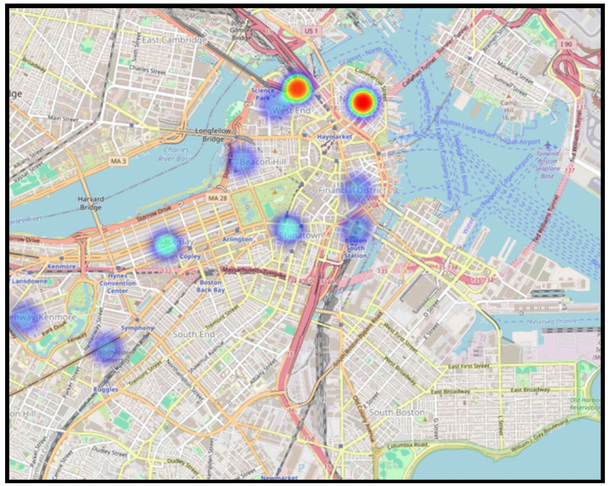
02

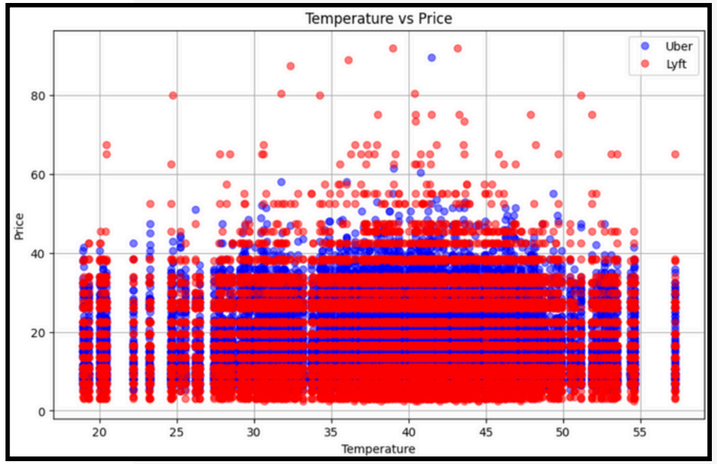
EDA

DATA VISUALIZATIONS

A selection of some of our visualizations







Price Histogram for Uber/Lyft

The **average price for Lyft is higher** at $17.42, whereas Uber has a mean of just $15.75. From the histogram, Lyft tends to have higher pricing, which was surprising to us, since Uber dominates the market share.

Folium Heatmap of Rides

Based on our ouput above, rather than an even distribution throughout Boston, we see certain areas with **concentrated frequencies** of ride sources. The most common areas are **Boston North Station and North End**, which might suggest higher prices here.

Temperature vs Price Scatterplot

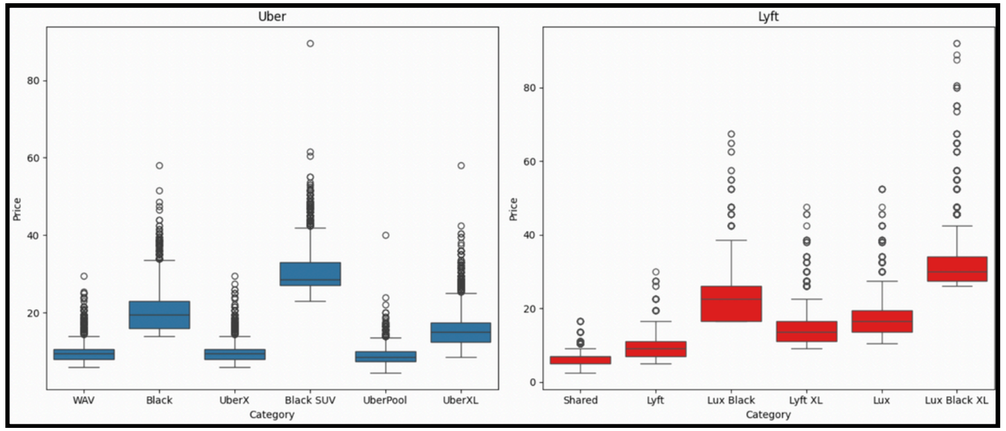
We see a **bell curve** type of trend, with peak prices around 35 degrees. This might indicate that in the **lower temperature** ranges there might have been **lower demand**, causing the companies to **lower their price**. However in higher demand times in the **middle temperature range**, rideshare companies probably had to **increase prices**. This would help mitigate the high demand by creating a **"scarcity" effect**.

02

EDA

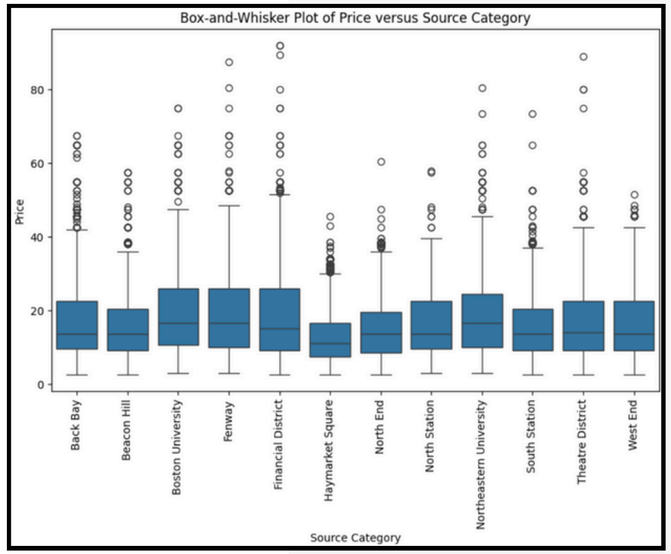
DATA VISUALIZATIONS

A selection of some of our visualizations



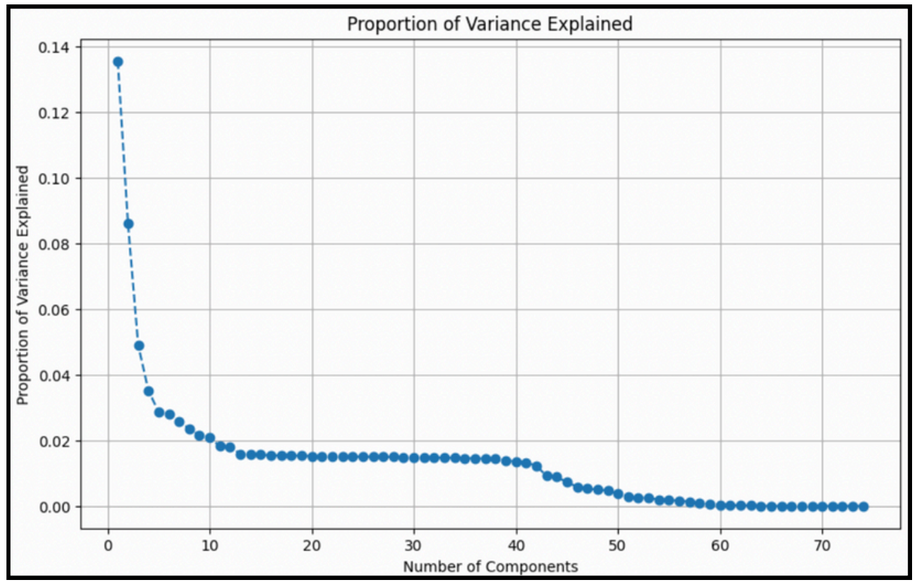
Box-and-Whisker plots of Ride Type versus Price

For Uber, the **Black SUV** rides are the **most expensive**, whereas for Lyft, the **Lux Black XL** rides are the most exponesive. Uber's **UberPool and Lyft's Shared** services tend to be the **chepeast**, which makes sense since that is a common way that ridesharing companies offer cheaper deals and **customer differentiation**: sharing a ride with other customers.



Box-and-Whisker plot of Price versus Area

The medians and variance are different for the different areas. For example, Financial District has very small spread compared to the other areas. Certain areas, like **North Station and Boston University**, have **higher medians**, aligning with our heatmap findings, whereas Beacon Hill, Back Bay, and Financial District have lower median prices.



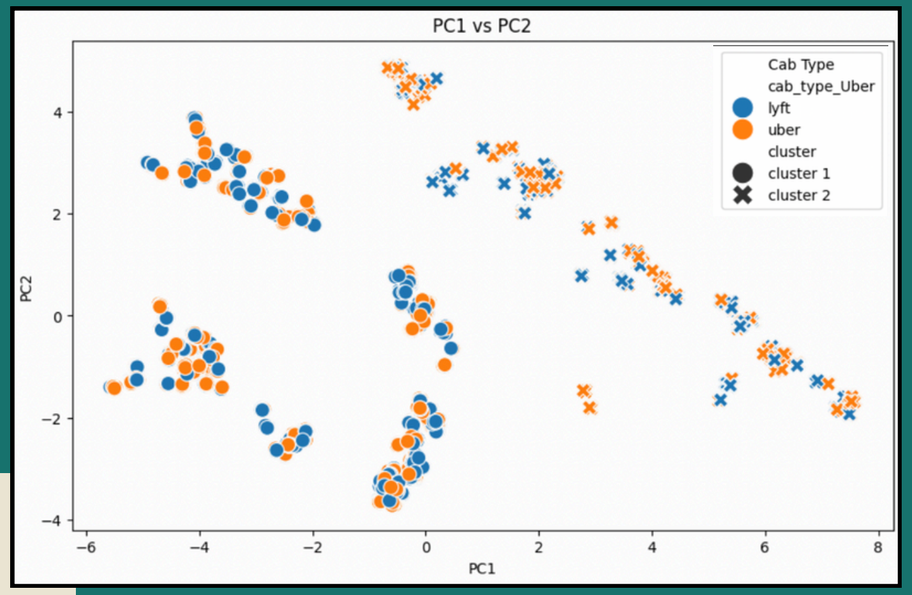
03 PCA & CLUSTERING

CLASSIFICATION OF UBER/LYFT RIDES

For our first model, we utilize PCA for **dimension reduction** by finding linear combinations of the indicator variables such that variance is maximized. By using a Proportion of Variance Explained line graph, we select the **top 5 PCAs** and perform clustering analysis to see if the model is able to discern between Uber and Lyft rides.

Downsides: PCs aren't easily understandable, simple model

By conducting this model, we can identify **if the rideshare companies are different and distinguishable**, or if Lyft has improved since its entrance in the market to have an even playing field against its competitor, Uber.



RESULTS

2D clustering plot:

- **colors** = company, **shape** = clusters
- **don't see a clear clustering effect**
- **misclassification error of 50.6%** (random chance) - Uber/Lyft **indistinguishable**

In terms of business analytics, over time, we can see that Lyft has risen to have very **similar characteristics** as Uber (dominant market share), becoming the second most popular rideshare app. This has allowed the company to price their services similar to Uber for rides that have similar features.

Drawbacks: **too simple**, **overlooking complexities in feature interactions,** the PCA variance might explain other differences instead of company

04

REGRESSION

REGRESSION MODELS FOR PRICE

Multiple Regression with No Penalty

We then aimed for **more interpretable results** than the PCA model by evaluating the beta and p-values for features in a multiple regression on the unscaled data. We printed the positive and negative beta value features separately.

Error Metrics

We found an **MAE of about 1.73 and 1.74** for train and test data respectively, and **MSE was 6.02 and 6.55** for train and test data respectively. These errors suggest the model **performs pretty well**, since the average price is about $16. Since the error is a bit higher for the test data set, this suggests this model is **slightly overfitted**. Due to the lack of penalties, this model might be **too complex.**

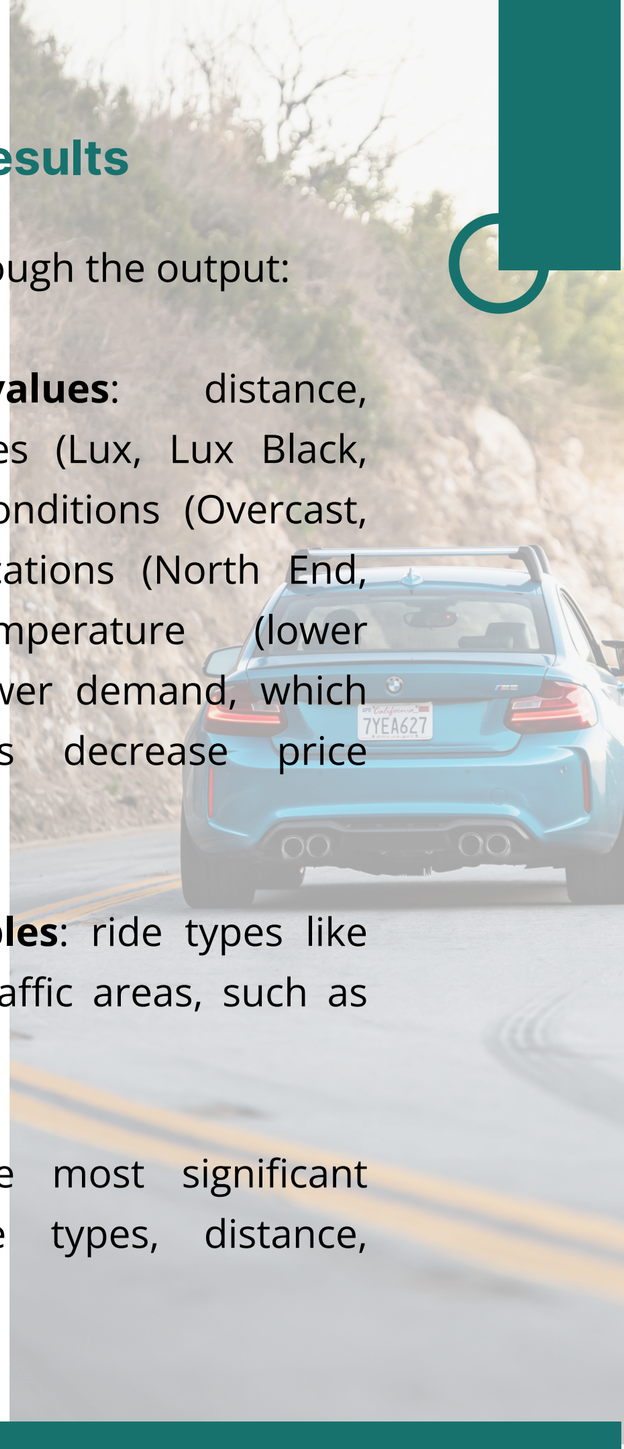
Interpretation of Results

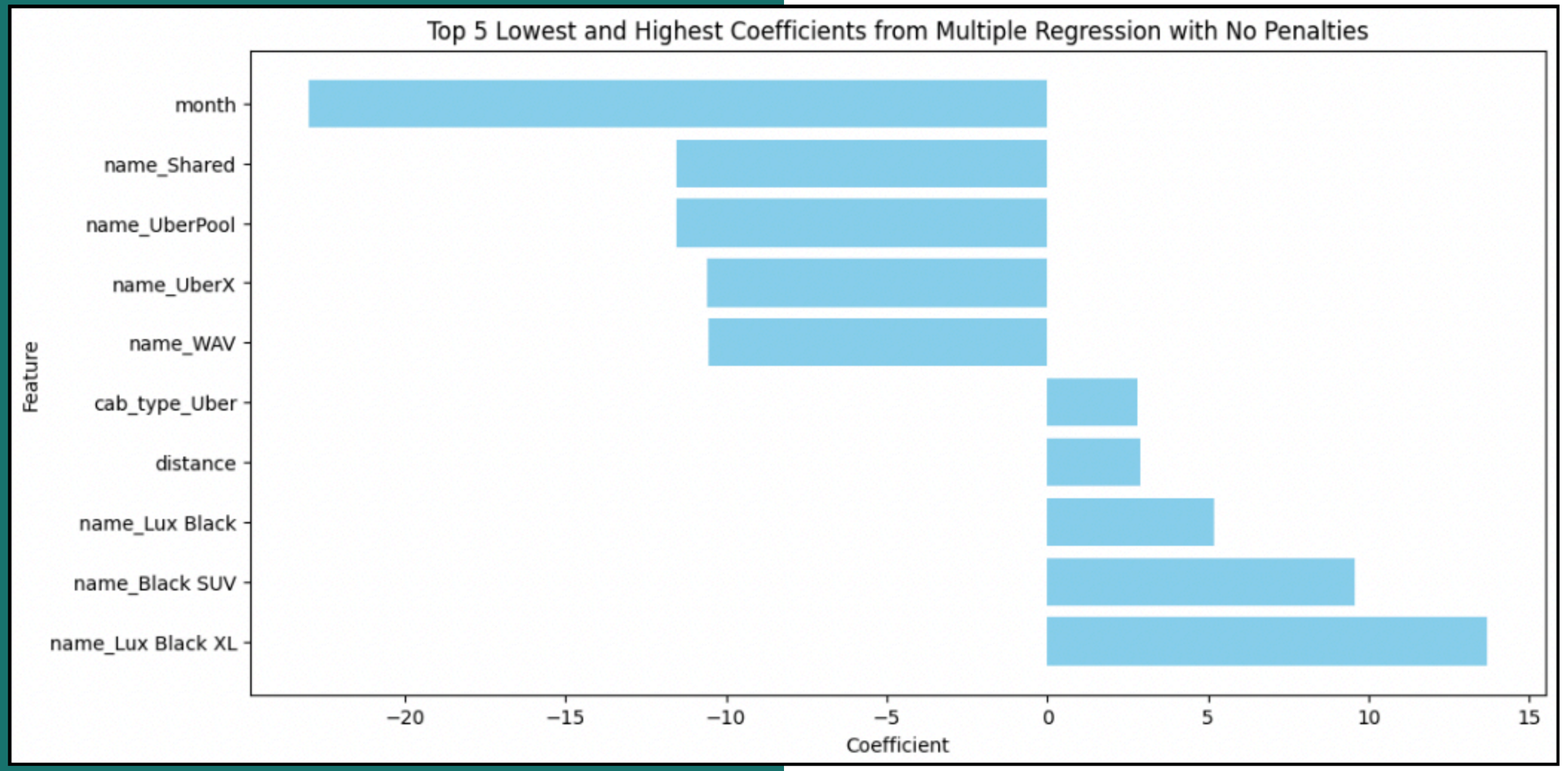
We used **PandaSQL** to sort through the output:

Positively correlated values: distance, precipitation, luxury ride types (Lux, Lux Black, Black SUV), worse weather conditions (Overcast, Drizzle, Rain), high traffic locations (North End, North Station), and temperature (lower temperatures might cause lower demand, which means rideshare companies decrease price accordingly).

Negatively correlated variables: ride types like UberPool/Shared and lower traffic areas, such as Beacon Hill and Fenway.

Low p-value variables: The most significant variables were special ride types, distance, Uber/Lyft, longitude, and day.




Top 5 Lowest and Highest Coefficients from Multiple Regression with No Penalties
month
name_Shared
name_UberPool
name_UberX
name_WAV
cab_type_Uber
distance
name_Lux Black
name_Black SUV
name_Lux Black XL
Feature
−20
−15
−10
−5
0
5
10
15
Coefficient

04 REGRESSION

REGRESSION MODELS FOR PRICE

Multiple Regression with Penalty

Since we saw overfitting in the regression with no penalty, we now aim to **reduce model complexity** through **regularization** (penalties). We use **Grid Search** for hyperparameter tuning and optimization for Ridge, LASSO, and Elastic Net models.

Error Metrics

Ridge yielded the best performance of the 3, with cross-validation scores of 0.92/0.91 for train/test, and an MSE of 6.01/6.55 & MAE of 1.73/1.74. These errors are almost identical to the regression w/o penalty, despite the reduced complexity. But it still shows **some overfitting**. Overall, this model **performed well**, giving similar results but with less variables. However, it overlooks variable interactions.

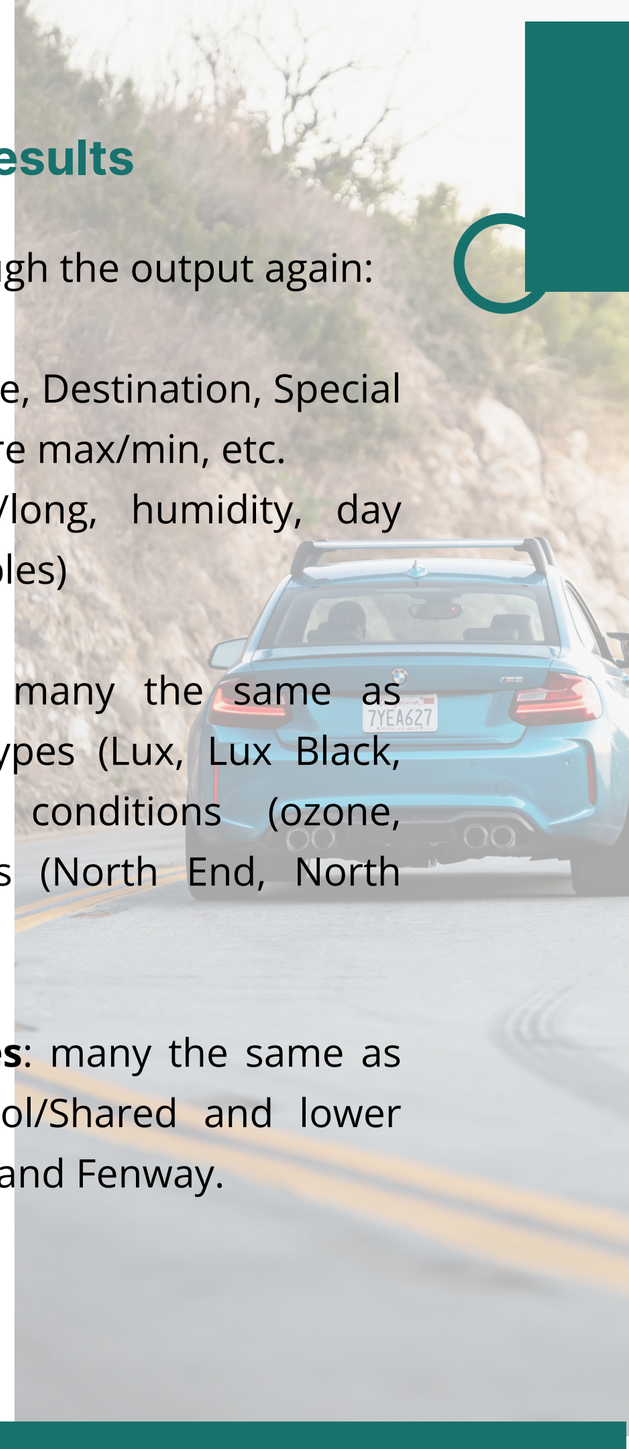
Interpretation of Results

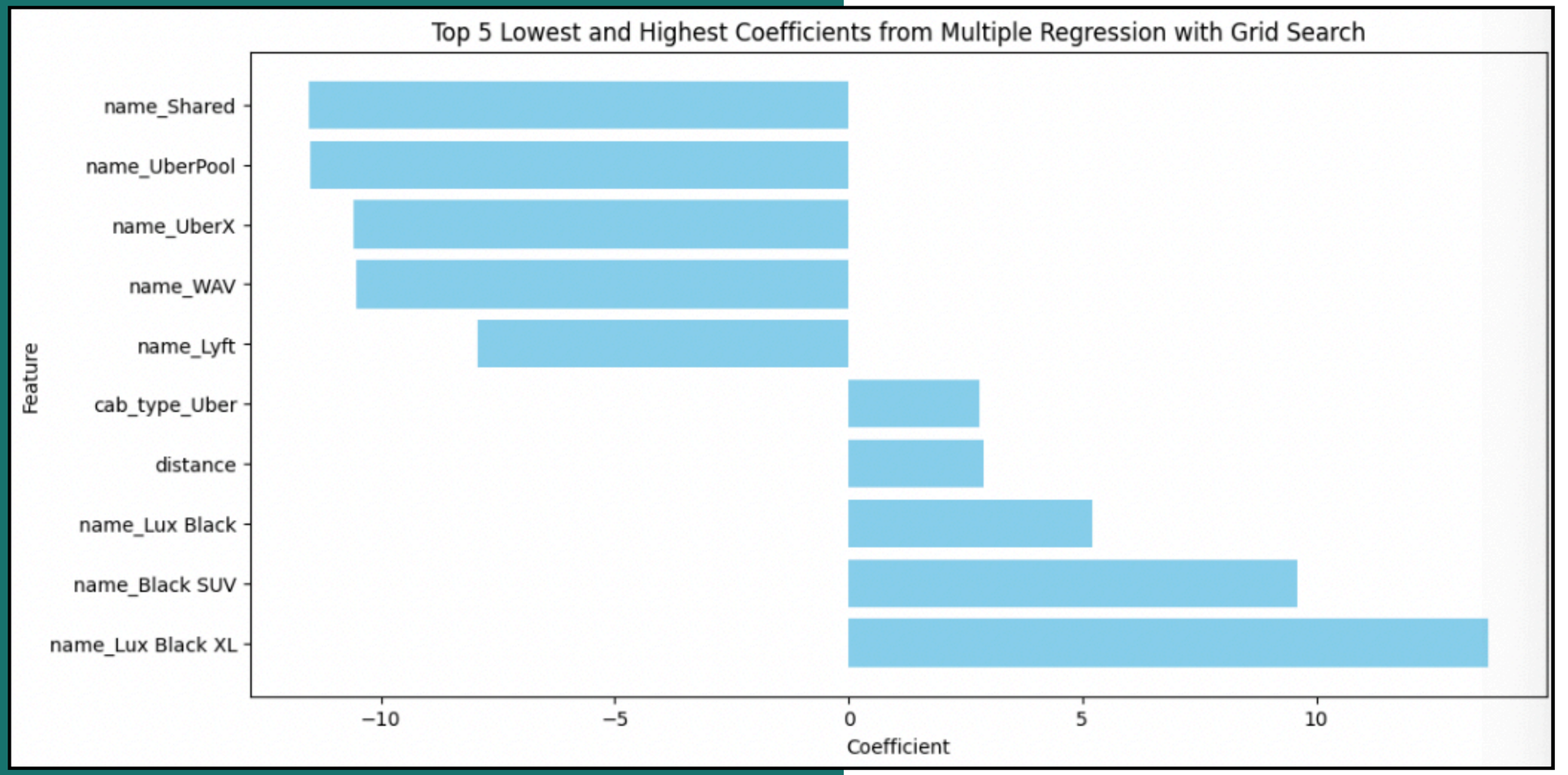
We used **PandaSQL** to sort through the output again:

Variables not kicked out: Source, Destination, Special Ride types, company, temperature max/min, etc.
Kicked Variables: month, lat/long, humidity, day length, etc. (many weather variables)

Positively correlated values: many the same as before: distance, special ride types (Lux, Lux Black, Black SUV), worse weather conditions (ozone, pressure), high traffic locations (North End, North Station), and temperature.

Negatively correlated variables: many the same as before: ride types like UberPool/Shared and lower traffic areas, such as Beacon Hill and Fenway.




Top 5 Lowest and Highest Coefficients from Multiple Regression with Grid Search
name_Shared
name_UberPool
name_UberX
name_WAV
name_Lyft
cab_type_Uber
distance
name_Lux Black
name_Black SUV
name_Lux Black XL
Feature
−10
−5
0
5
10
Coefficient

05 NEURAL NET

FEEDFORWARD NEURAL NETWORK

```
Epoch 1/10, Loss: 1.00125473122508
Epoch 2/10, Loss: 1.0008021588051177
Epoch 3/10, Loss: 1.0012827957342119
Epoch 4/10, Loss: 1.0007302272945087
Epoch 5/10, Loss: 1.0004233226759986
Epoch 6/10, Loss: 1.0008740611689546
Epoch 7/10, Loss: 1.0011550854708742
Epoch 8/10, Loss: 1.0004175660372183
Epoch 9/10, Loss: 1.000687440639825
Epoch 10/10, Loss: 1.0001628571355403
```

```
Test Loss: 1.0092180733169829
```

01 Model to predict Price

Since our previous models don't investigate variable interactions as much, we now use an FNN on the scaled data to do so. The architecture uses **4 linear layers** and **ReLU activation**. The last layer has 1 out feature since we're predicting one variable (price).

02 Model Training

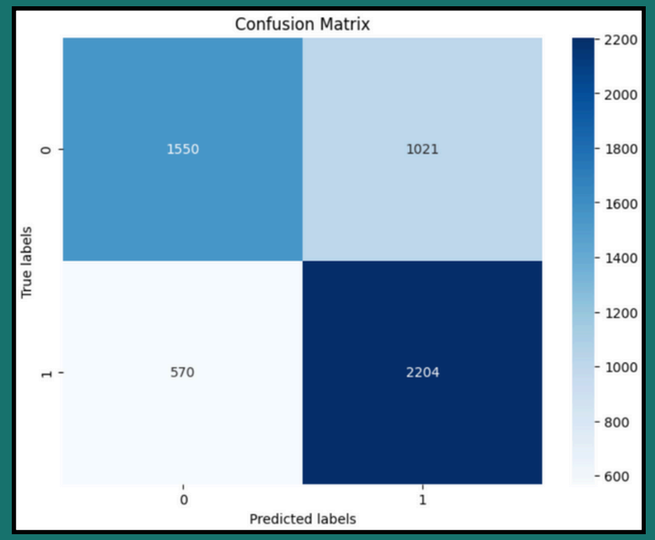
We use MSE loss and an **Adam optimizer** along with 10 epochs to train the model. The Adam optimizer optimizes our parameters in the neural network.

03 Performance Analysis

We see the testing error **(MSE) to be 1.01**, which is pretty high for scaled data. This means that our model performed relatively worse compared to our previous regression models. We can also see that the model is slightly overfitted. This performance might be low because there might be too much noise in the dataset and the model has trouble identifying patterns.

06 RANDOM FOREST

CLASSIFYING UBER AND LYFT RIDES



Training Accuracy:
0.827
Training Recall:
0.904
Training Precision:
0.792

Testing Accuracy:
0.702
Testing Recall:
0.795
Testing Precision:
0.683

Model Justification

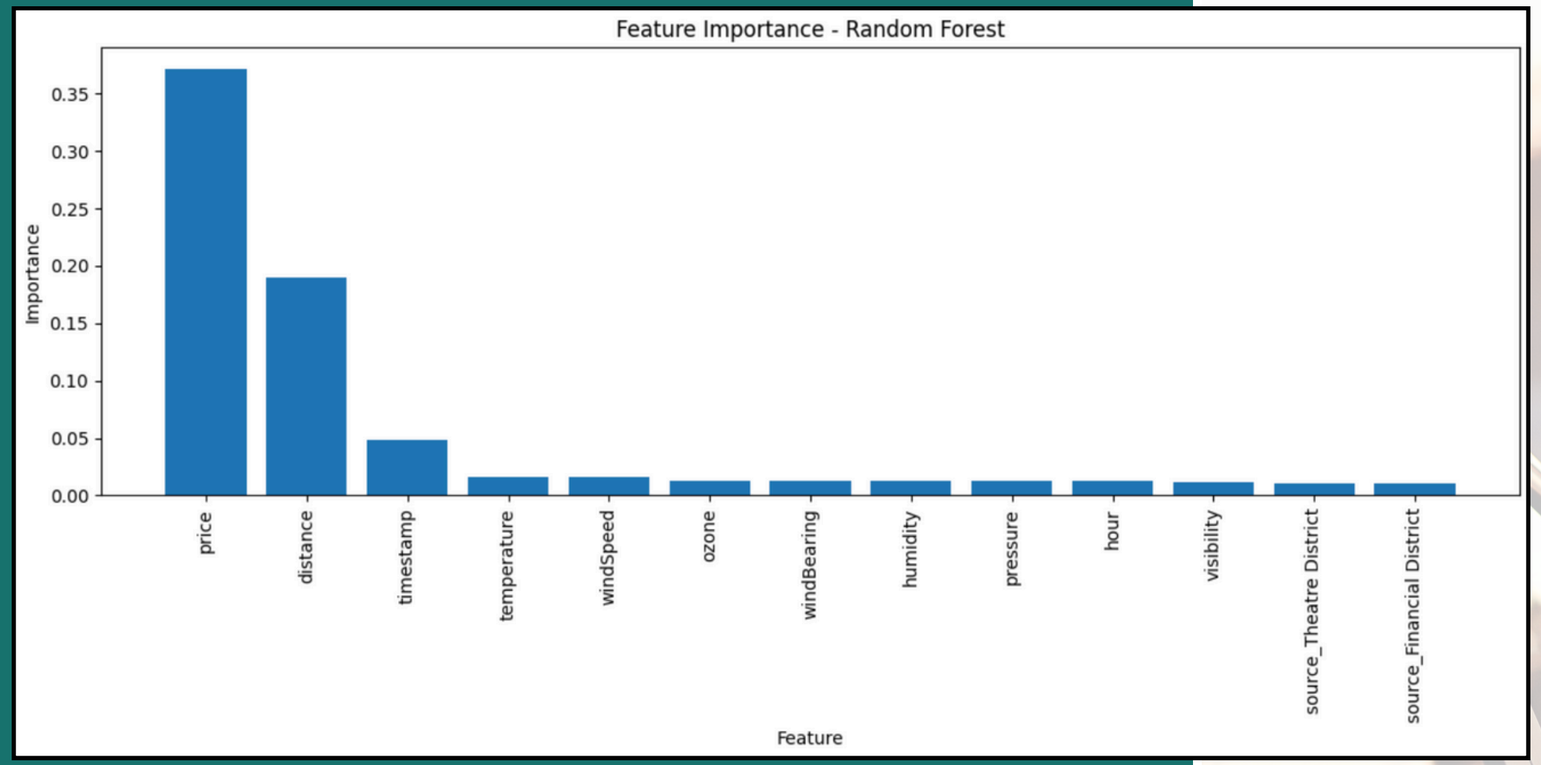
Going back to our **classification** problem now, we want to build a more **complex** model than the simple clustering from earlier, and one that is more **interpretable** than PCs. Thus we use a more powerful model: Random Forest, **removing special ride types** (specific to Uber/Lyft)

Hyperparameter Tuning and Training

We use 3-fold cross validation **grid search** to optimize our **hyperparameters**, including n_estimators, max_depth, min_samples_split, and min_samples_leaf.

Error Analysis

Training accuracy of 82.7% with a recall of 90.4% and a precision of 79.2%, indicating that it is quite effective at identifying the correct class labels during training (high sensitivity), particularly in correctly predicting the **majority class**. However, in testing, **accuracy** dropped to **70.2%**, with a **recall of 79.5%** and a precision of 68.3%, suggesting the model might be **overfitted** and thus performing less effectively on unseen data. Despite that, this model is much **more accurate than our clustering**.



Feature Importance for Random Forest

As we can see from the **feature importance** chart, the most important features for reducing entropy in the random forest were **price** and **distance**. This suggests that these are the features that distinguish Uber from Lyft the most.

07 RESULTS & INSIGHTS

Regression results: Most important factors for price were high traffic areas (North Station, North End), distance, special ride types (UberXL, Lux), and worse weather conditions (rain, overcast). **Classification results**: Our PCA/clustering model cannot distinguish b/w Uber/Lyft, suggesting Lyft has implemented good business tactics as a competitor. However, our random forest was better at classifying rides by rideshare company. Based on our EDA, we found that Lyft actually tends to have higher prices than Uber.

Based on our various models, we saw the best classification performance from the random forest model (80% recall, 70% accuracy) and the best regression performance from the Ridge regression model (MSE: 1.74). While some had overfitting, adjusting complexity helped hone in on key variables involved and improve interpretability.

CONSUMERS	**Customers, such as college students, can use these insights to better predict when prices might be high.**
BUSINESSES	**Other emerging rideshare companies can look to Lyft's successful business strategies to become a competitive player and be able to raise prices.**

08 CONCLUSIONS & DISCUSSION

Challenges & Limitations

Our dataset was limited to just the Boston area, so it only picked up on **urban rideshare patterns**, and only had data from a **limited 2 month** timeframe. Exploring data from when **Lyft had just entered** the market would have more differentiation. It also only contained two of **many rideshare companies**.

Future Work & Models

For further analysis, a **Boosting** model might have improved accuracy and reduced bias due to its creation of strong learners for classifying rideshare companies. We could also try different **achitectures** for the **neural network** for optimization. Lastly, we could explore **time series models.**

Other Data

Explore datasets that included a **wider timeframe** (rather than just two months), and a wider **geographic region** (rather than just Boston). Since Boston is an urban location, analyzing more **diverse data**, would increase differentiation to improve our models' predictions and insights.

THANK YOU!

May 1 2024