UBER & LYFT RIDE ANALYSIS

JOAN CONDRON, DINA ELOSEILY, SWATHI PRAKASH

01

Background

OBJECTIVE

Analyze Uber and Lyft ride hailing data to identify key patterns and factors contributing to trip prices.

VALUE PROPOSITION

1. Pricing prediction models provide ridesharing companies the opportunity to ensure their pricing algorithm is working as expected.

2. Predictive models could allow companies to more accurately forecast demand and revenue.

3. This analysis could help better understand the factors that contribute to ride prices.

DATASET

Name: Uber and Lyft Dataset Boston, MA
Source: Kaggle

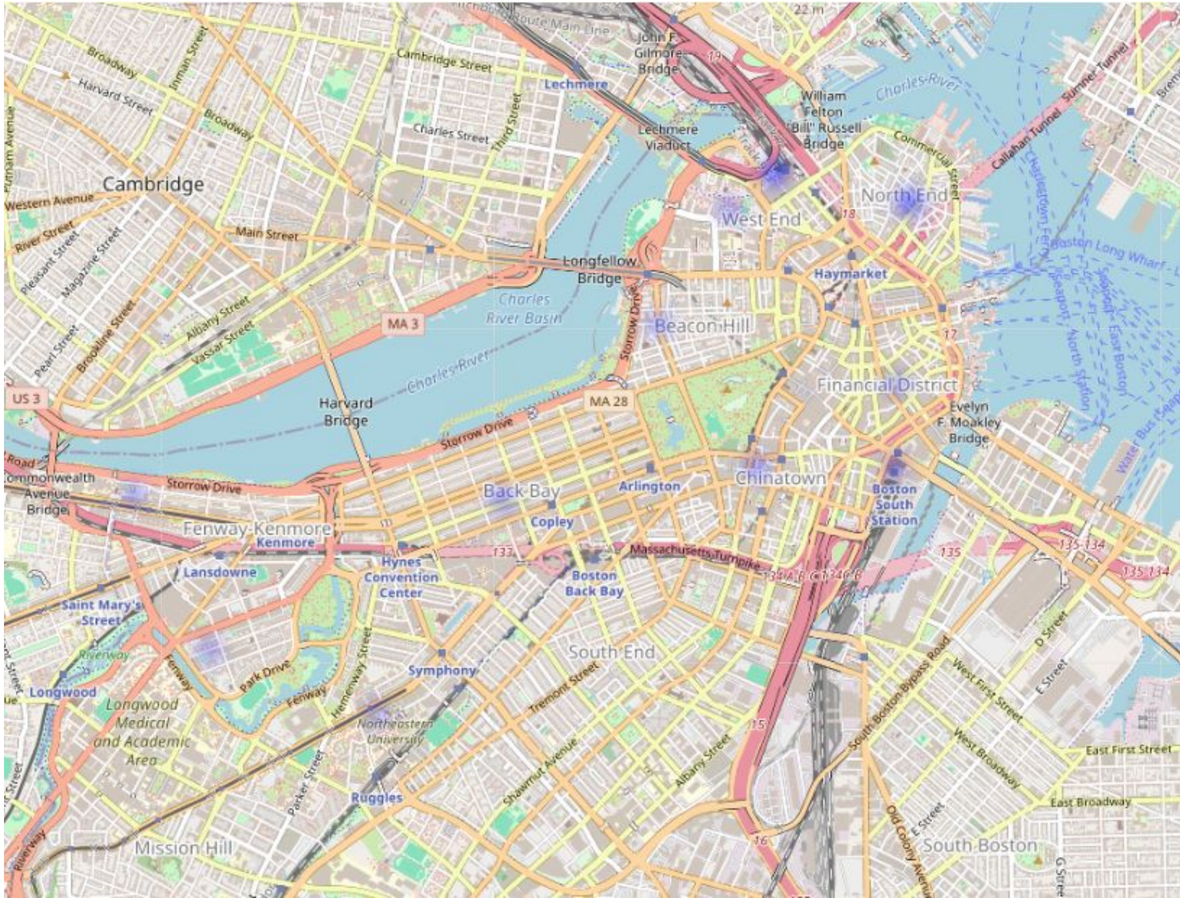
- 693k rows: records capture ride information over two months from November to December 2018 in the Boston area
- 57 Columns: contains price, date, time, destination, ride type, and various weather features

02

EDA

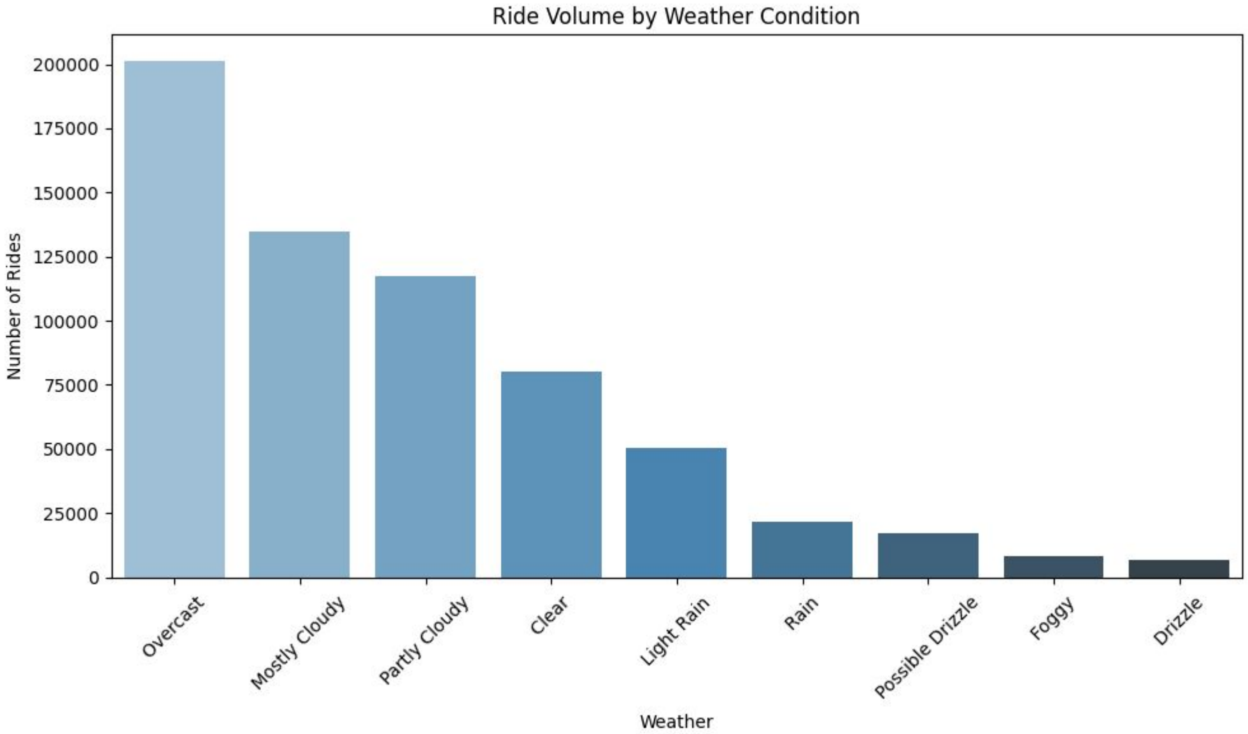
Ride Density

Geomap visualization showing the distribution of Uber and Lyft rides. Each spot represents a destination, with the shading indicating the volume of rides—darker spots correspond to higher ride activity.



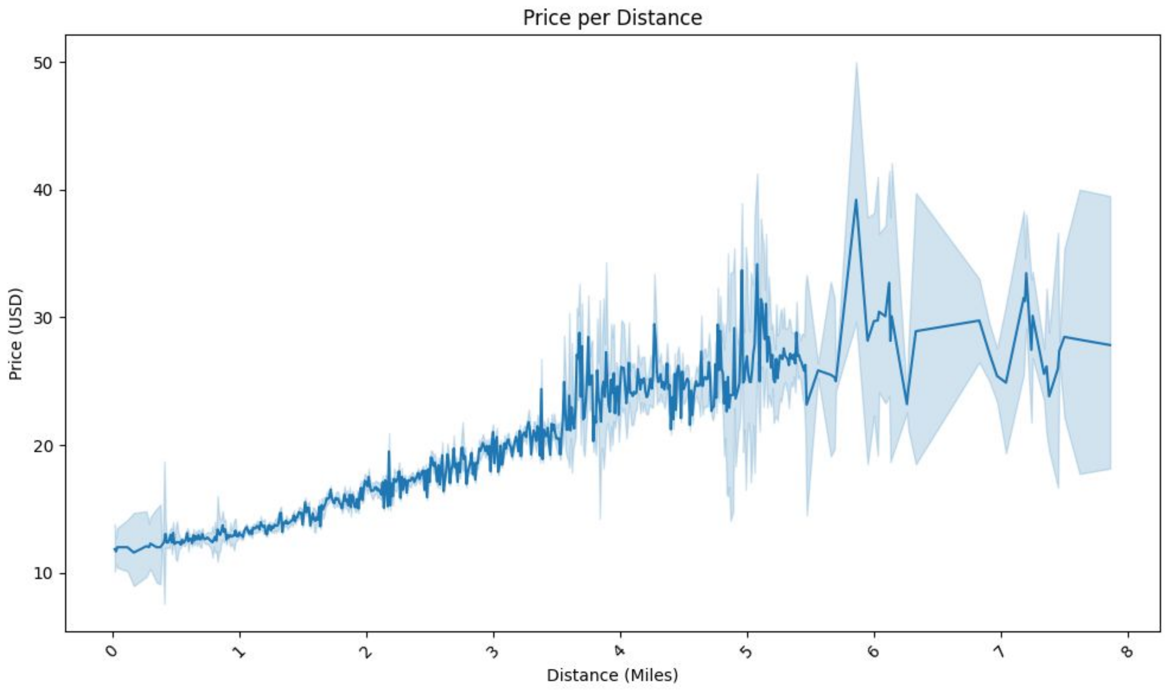
Impact of Weather on Ride Volume

Overcast, Mostly Cloudy, and Partly Cloudy days saw the highest number of Uber and Lyft rides, while Drizzling conditions had the lowest ride volume.



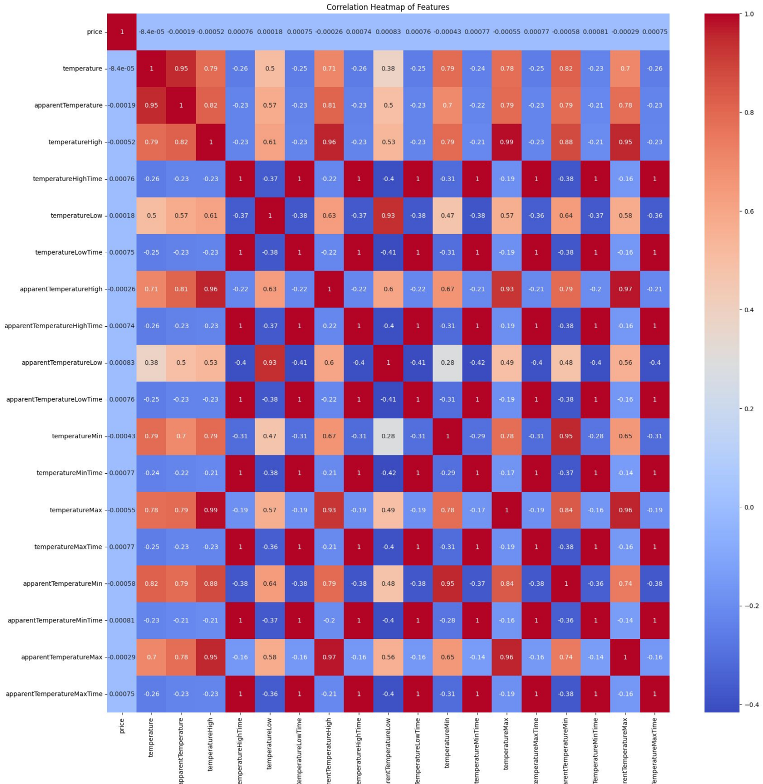
Impact of Distance on Ride Price

Preliminary visual analysis suggests there may be a relationship between price and distance of rides.



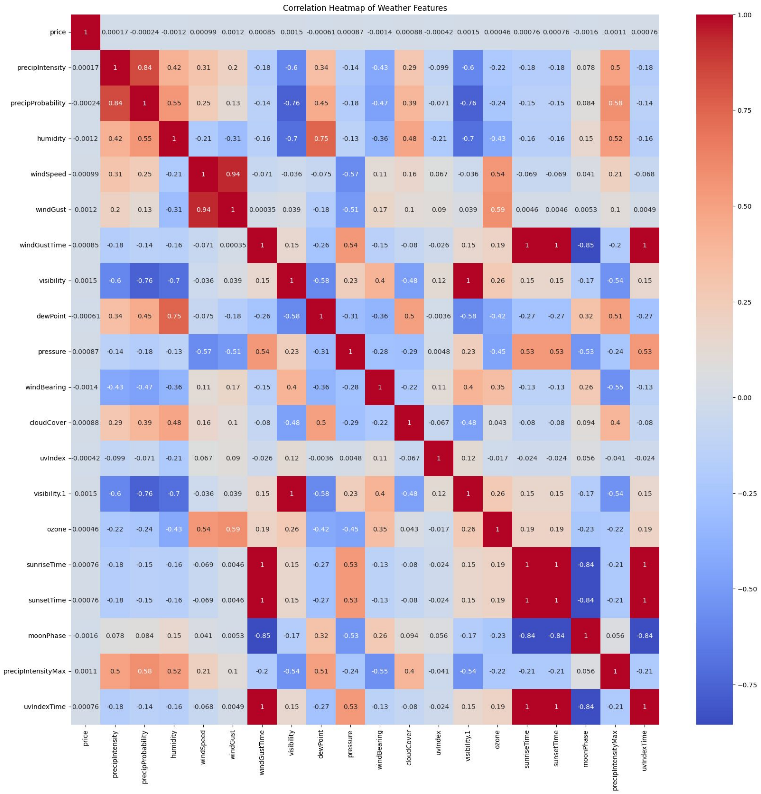
Correlation Heat Map: Price & Temperature Variables

Takeaway: Temperature has minimal influence on pricing dynamics



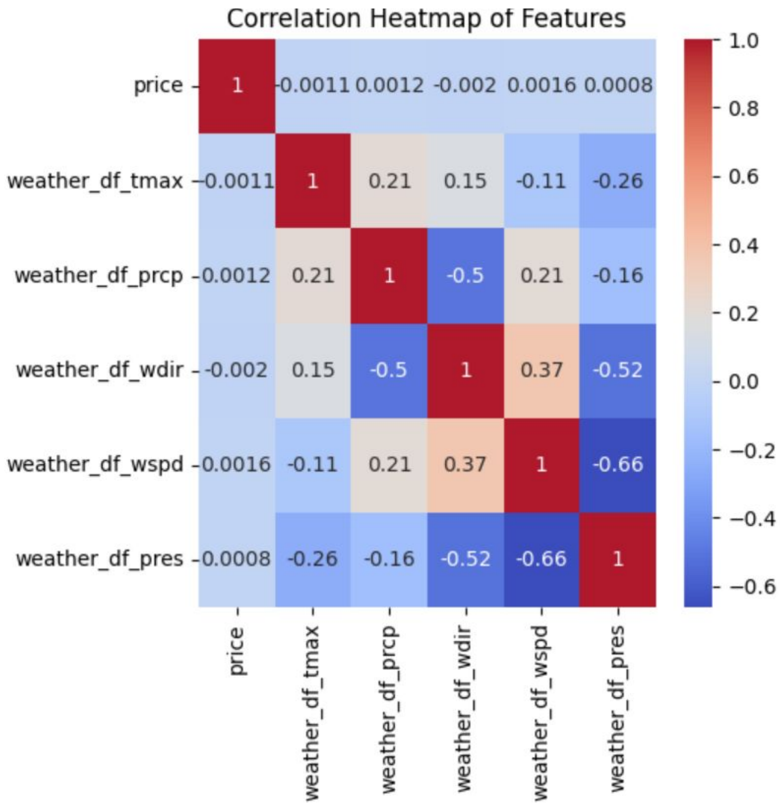
Correlation Heat Map: Price & Other Weather Variables

Takeaway: Weather related features have minimal influence on pricing dynamics



Validating weather's impact with external data

Takeaway: weather has little effect on pricing.



03

MODELING

Modeling Overview

1. Lasso (L1) Linear Regression
2. Ridge (L2) Linear Regression
3. Random Forest Regression
4. XGBoost Regression

Lasso (L1) Linear Regression

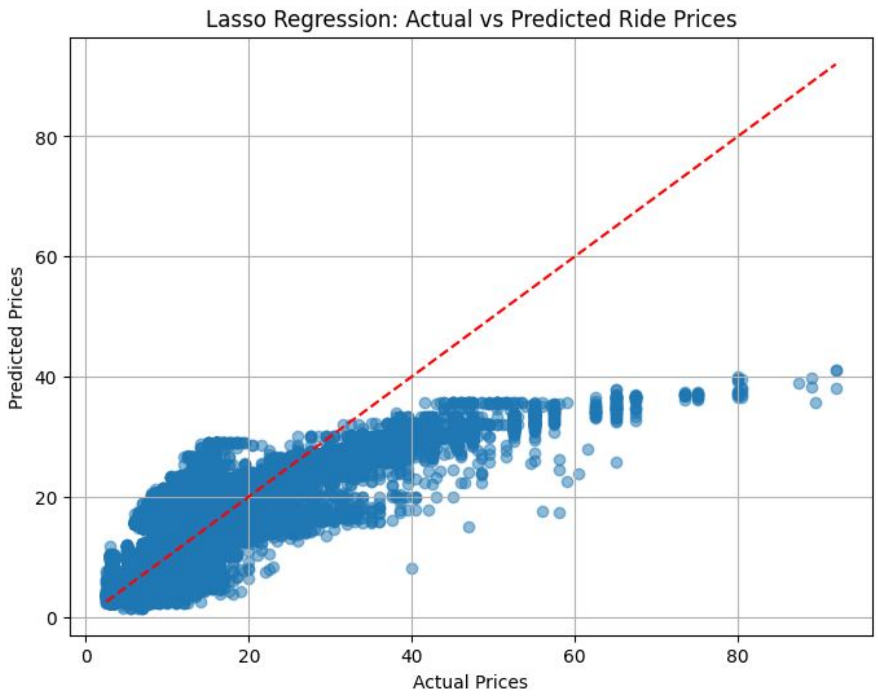
MSE: 22.55

$R^2_{training}$ = 0.7424

R^2_{test} = 0.7413

Features with non-zero Coefficients:

Distance
Surge Multiplier
Product Name



Ridge (L2) Linear Regression

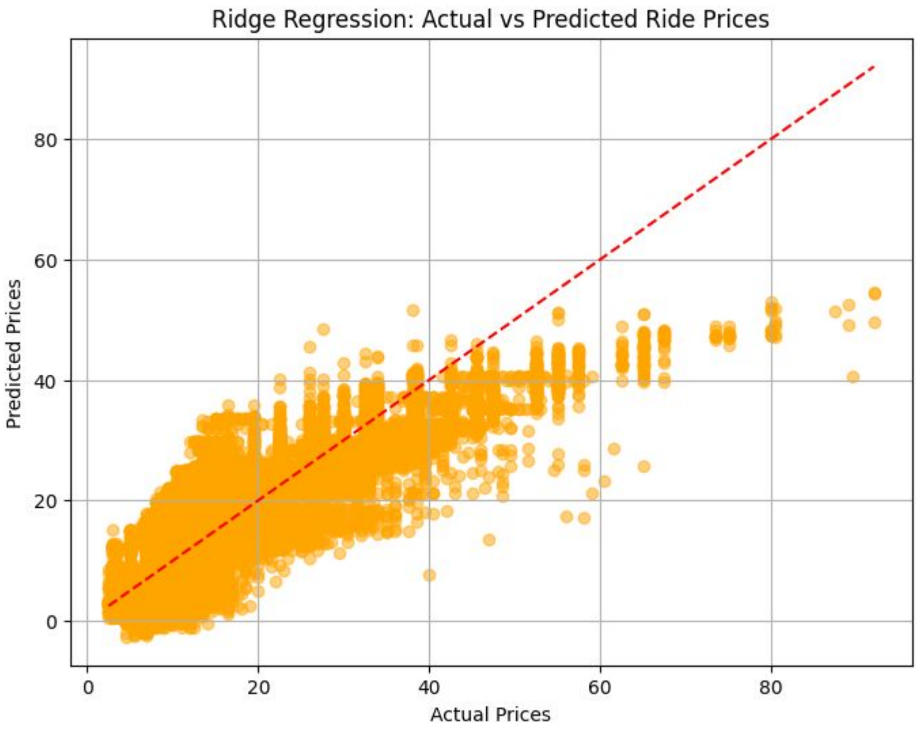
MSE: 19.60

$R^2_{training}$ = 0.7755

R^2_{test} = 0.7752

Selected Features:

- Distance
- Surge Multiplier
- Product Name
- Starting Destinations
- End Destinations



Random Forest Regression

Grid Search
Number of Estimators= [100, 150] **Max Depth= [3, 5]**

Best Model
Estimators= 150
Max Depth= 5

Performance
MSE: 6.66 $R^2_{training}$ = 0.9232 R^2_{test} = 0.9236 **Important Features:** Product Name Distance Surge Multiplier Cab Type

XGBoost Regression

Grid Search
Learning Rate = 0.01 **Number of Estimators= [50, 100, 150]** **Max Depth= [3, 5]**

Best Model
Estimators= 150
Max Depth= 5

Performance
MSE: 2.82 **$R^2_{training}$=** 0.9673 **R^2_{test}=** 0.9676 **Important Features:** Product Name Distance Surge Multiplier Cab Type

03

RESULTS

Summary

Model	MSE Loss	R^2 Train	R^2 Test
Lasso	22.5	0.7424	0.7413
Ridge	19.6	0.7752	0.7755
Random Forest	6.57	0.9232	0.9236
XGBoost	2.82	0.9673	0.9676

- Performance Ranking:
  - XGBoost
  - Random Forest
  - Ridge
  - Lasso
- Ensemble methods performed better because they can model non-linear relationships

04

CONCLUSIONS

Key Takeaways

1. **Main features** influencing price:
   a. Product Type (UberPool vs UberXL)
   b. Distance
   c. Surge Multiplier
   d. Cab Type (Uber vs. Lyft)
2. XGBoost and Random Forest have the ability to model non-linear relationships which improved the performance compared to the linear models
3. **XGBoost** model had the best performance
4. Training for **XGBoost was significantly faster** (~4 mins) for 30 fits rather than Random Forest (~44 mins) for 20 fits

Limitations

1. Model trained on data from November-December 2018
    a. OK- if looking at specific time periods, e.g prior month, one quarter
    b. More generalized, train on multi-year data
2. Data only contains rides in Boston, MA
    a. May not be representative of other locations (eg. suburbs, west coast)
3. Data is missing information on events occurring in Boston which could influence demand pricing
4. There is no information on pricing promotions

Future Considerations

1. Expand the data date range to multi-year
   a. This would really make it Big Data!
2. Join user data to explore how consumer’s ride frequency, passenger rating, or total spend impact the price
3. Explore stacking where the Random Forest and XGBoost are the base learners
4. Instead of Grid Search, implement hyperparameter tuning with Ax

Modeling Implications for Stakeholders

1. **Verify** the dynamic pricing algorithm is working as expected
2. Test different **pricing strategies** or promotional offerings
3. **Analyze bias** and fairness and to ensure there are no discriminatory pricing practices
4. Consumers can make more **informed** ride hailing decisions

THANK YOU