Data Science - Assignment 2
Ride Sharing Analysis - Uber & Lyft

Mohammad Hossein Basouli

May 7, 2025

Linear & Polynomial Models

Data Preprocessing

Data Reduction: We remove irrelavant features as well as redundant features. e.g. 'id', 'timestamp', 'timezone', 'datetime', 'visibility.1', 'temperatureMinTime', 'temperatureMax', 'temperatureMaxTime', 'apparentTemperatureMin', 'apparentTemperatureMinTime', 'apparentTemperatureMax', 'apparentTemperatureMaxTime'

Data Transformation:

- extraction of **sunriseHour** and **sunsetHour**.
- encoding of categorical features via **One-Hot Encoding**.

Exploratory Data Analysis

High Correlated Features: Features with a correlation absolute value greater than 0.05:

Feature	Value
price	1.000000
distance	0.333871
surge_multiplier	0.165611
source_Boston University	0.067880
source_Fenway	0.057167
source_Haymarket Square	-0.101514
destination_Boston University	0.072743
destination_Fenway	0.053153
destination_Haymarket Square	-0.075808
destination_South Station	-0.055659
cab_type_Uber	-0.068606
product_id_6c84fd89-3f11-4782-9b50-97c468b19529	0.195100
product_id_6d318bcc-22a3-4af6-bddd-b409bfce1546	0.412593
product_id_997acbb5-e102-41e1-b155-9df7de0a73f2	-0.300286
product_id_9a0e7b09-b92b-4c41-9779-2ad22b4d779d	-0.236477
product_id_lyft	-0.239700
product_id_lyft_line	-0.426482
product_id_lyft_lux	0.245791
product_id_lyft_luxsuv	0.412894
product_id_lyft_premier	0.107153
name_Black SUV	0.412593
name_Lux	0.107153
name_Lux Black	0.245791
name_Lux Black XL	0.412894
name_Lyft	-0.239700
name_Shared	-0.426482
name_UberPool	-0.300286
name_UberX	-0.236468
name_WAV	-0.236477

Table 1: Features with a correlation absolute value greater than 0.05

Feature Selection: At the begining I have utilized **sklearn univariate feature selection** methods such as **VarianceThreshold**, **SelectKBest** and **SelectPercentile** along with metrics such as **r_regression**, **f_regression** and **mutual_info_regression**, but they all gave poor results. It might be due to this fact that these methods might not work in some cases, e.g. for a dataset which even the low correlated features lead to a significant, important information in our analysis. This might be the case, since we haven't got good results .Thus, we approach the **Feature Selection** process by an-

other method, like examining the relation of correlated features with the target variable by hand. It turns out that this approach leads to very decent results, thus, we will only maintain features that are listed in the table 1.
Outliers: First, we have removed the outliers from our dataset but it did no good. Also the dataset size is very large, even if the dataset contains outliers, this could be ignored because our sample size is enough in order to make the model robus to the outliers. Thus we keep all of the rows.

Train-Test Splitting

20% Test, 80% Train

Linear Models

Base Linear Regression With Missing price Dropped

Evaluation:

Metric	**Train**	**Test**
MSE	6.2973	6.3036
MAE	1.7688	1.7684
R^2	0.9275	0.9277

Table 2: Evaluation of Base Linear Regression With Missing **price** Dropped

Coefficients:

Feature	Coefficient
distance	3.248531
product_id_lyft_luxsuv	2.314652
name_Lux Black XL	2.314652
name_Black SUV	2.051601
product_id_6d318bcc-22a3-4af6-bddd-b409bfce1546	2.051601
surge_multiplier	1.750053
name_UberX	-1.659294
product_id_6c84fd89-3f11-4782-9b50-97c468b19529	1.358077
name_Shared	-1.165846
product_id_lyft_line	-1.165846
product_id_lyft_lux	1.053897
name_Lux Black	1.053897
name_UberPool	-0.973288
product_id_997acbb5-e102-41e1-b155-9df7de0a73f2	-0.973288
product_id_9a0e7b09-b92b-4c41-9779-2ad22b4d779d	-0.829638
name_WAV	-0.829638
name_Lyft	-0.776932
product_id_lyft	-0.776932
cab_type_Uber	0.528733
product_id_lyft_premier	0.336811

Table 3: Top features with corresponding coefficients

Residual Plot:

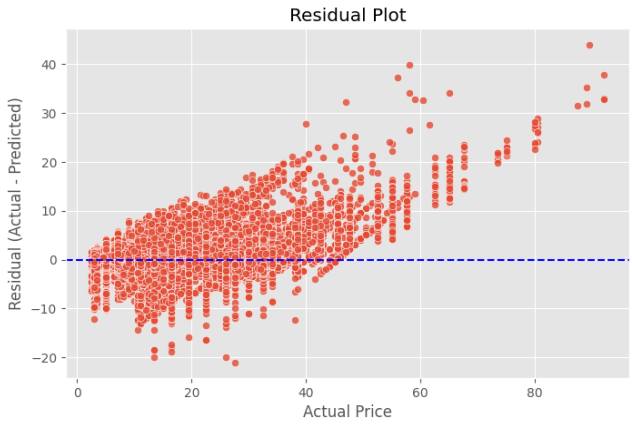

Figure 1: Residual Plot of Base Linear Regression With Missing **price** dropped.

Base Linear Regression With Missing price Returned Back to The Training Set

Evaluation:

Metric	Train	Test
MSE	32.8414	32.8825
MAE	3.1355	3.1392
R^2	0.6220	0.6228

Table 4: Evaluation of Base Linear Regression With Missing **price** Returned Back to The Training Set

Regularized Linear Ridge Regression

Evaluation:

Metric	Train	Test
MSE	6.2973	6.3036
MAE	1.7688	1.7684
R^2	0.9275	0.9277

Table 5: Evaluation of Regularized Linear Ridge Regression

Regularized Linear Lasso Regression

Evaluation:

Metric	Train	Test
MSE	6.2973	6.3037
MAE	1.7683	1.7679
R^2	0.9275	0.9277

Table 6: Evaluation of Regularized Linear Lasso Regression

Polynomial Models

Quadratic Model With Ridge Regression

Best Alpha: 10
Evaluation:

Metric	Train	Test
MSE	3.4229	3.3842
MAE	1.2621	1.2582
R^2	0.9606	0.9612

Table 7: Evaluation of Quadratic Model With Ridge Regression

Quadratic Model With Lasso Regression

Best Alpha: 0.001
Evaluation:

Metric	Train	Test
MSE	3.4232	3.3843
MAE	1.2621	1.2582
R^2	0.9606	0.9612

Table 8: Evaluation of Quadratic Model With Lasso Regression

Enhanced Model

Data Preprocessing

Data Reduction: We remove irrelavant & redundant features. e.g. 'id', 'timestamp', 'timezone', 'datetime', 'visibility.1', 'temperatureMinTime', 'temperatureMax', 'temperatureMaxTime', 'apparentTemperatureMin', 'apparentTemperatureMinTime', 'apparentTemperatureMax', 'apparentTemperatureMaxTime'

Data Transformation:

- encoding of nominal categorical features via **One-Hot Encoding**
- encoding of ordinal categorical features via **Label Encoding**. This could improve our results since maintains difference in order of the labels in the column.
- extract **sunriseHour** and **sunsetHour**

Missing Values: We will drop the rows which have missing value on **price**.

Train-Test Splitting

20% Test, 80% Train

Evaluation

Metric	Train	Test
MSE	3.0890	3.0483
MAE	1.1793	1.1751
R^2	0.9644	0.9650

Table 9: Model performance metrics on training and test sets

Residual Plot

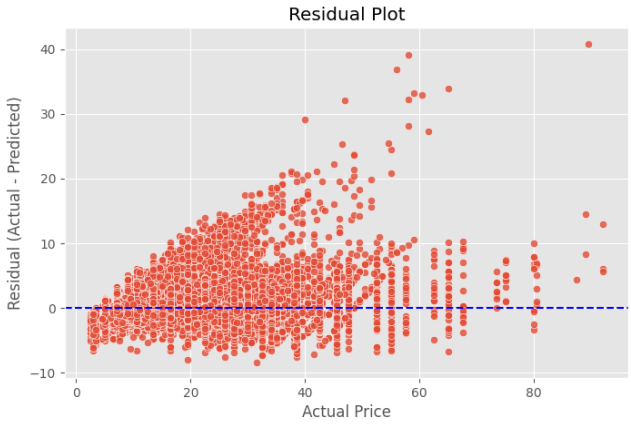

Figure 2: Residual Plot of Enhanced Model

Conclusion

Best Model: The best model is **Enhanced Model** which has the lowest **MSE**, **MAE** and also has the highest R^2 score which means the model explain much of the variance that exists whithin the data (both in **Train** and **Test** evaluation). This could be due to better encoding of ordinal features, e.g. **name** feature, as well as using good regularization regressions such as **Ridge Regression** which helps to avoid overfitting.