CAPITAL BIKESHARE REPORT

By: Sahith, Kason Richard, Shubham Patil, Nour Alzaid

Abstract:
This report conducts an exploratory analysis and predictive modeling of ride-hailing service data to assess the influence of weather conditions on pickup and drop-off counts. Various regression models including Linear Regression, Ridge Regression, LASSO, Elastic Net, and KNN Regressor were assessed for their effectiveness in predicting counts based on temperature, precipitation, and humidity as features.

Introduction:
The following report provides a thorough examination of how weather conditions influence the utilization trends of the Capital Bikeshare system, focusing on the "22nd & H St NW" station during the span of six months from January to June 2023. Our aim was to employ predictive models to enhance the distribution of bikes and docks in response to diverse weather conditions. We employed Linear Regression, Ridge Regression, LASSO, Elastic Net, and KNN Regressor models to forecast the number of pickups and drop-offs based on temperature, precipitation, and humidity as predictor variables.

Exploratory Analysis
As a group we were tasked with preparing code and analysis in Jupyter Notebook on bikeshare system data from Capital Bikeshare and weather data. The data spans from February 2023 to June 2023 at the dock location at 22nd & H St NW. We want to explore the relationships among the weather features like Temperature, Windspeed, Dew, Humidity, Precipitation, Feelslike, UVindex and the target variables pu_ct for Pick-up Count and do_ct for Drop-off Count utilizing visualizations like a scatterplot matrix, heatmaps, pairplots, etc. We did this analysis by downloading the data from csv files, grouping and manipulating the data to get pick-up and drop-off occurrences, and merge the weather data with the pick-ups and drop-offs data. We created a correlation matrix for the pick-ups data and another correlation matrix for the drop-offs data to find what features were reasonable in predicting. After performing Correlation between Pickups and drop-offs, we found that temp, precip, solarenergy, uvindex, feelslike and humidity are highly correlated. So when we have performed the Linear Regression using OLS, we found solar energy and humidity are having p-value greater than 0.05, which makes them insignificant. Then with the remaining independent variables we have continued our analysis. We further did the analysis by building different regression models with training and testing data for each feature individually while adding additional features one at a time. We then observe and determine the best models for predicting based on the Mean Squared Error (MSE) and then evaluate them.

Predictive Modeling:
Now since we need to predict the future outcomes like bike pickups and drop offs we are using predictive modeling techniques.
Predictive modeling is a process used in data analysis to predict future outcomes based on historical data and existing trends. It involves building a statistical model or machine learning algorithm that learns from past data to make predictions about future events or behaviors.

Linear Regression

In our Linear Regression model, we used the weather data Temperature, Precipitation, UVindex, and Feelslike. We split the data into a training and testing set and scaled the features for pick-ups and drop-offs.

For Pickups:

```
The coefficients are:
temp        -29.103308
precip       -2.607130
uvindex       2.154027
feelslike    32.372010
dtype: float64
```

For drop-offs:

```
The coefficients are:
temp        -26.399059
precip       -2.755314
uvindex       2.125999
feelslike    29.800688
dtype: float64
```

Prediction Performance:
The mean squared error (MSE) for the training set for drop-offs is approximately 63.97. This indicates the average squared difference between the actual drop-off locations and the predictions made by the model on the training data. The MSE for the testing set for drop-offs is approximately 84.15. This indicates the average squared difference between the actual drop-off locations and the predictions made by the model on the testing data. A higher MSE suggests that the model's predictions are less accurate on the testing data compared to the training data. The MSE for the training set for pick-ups is approximately 64.98. This indicates the average squared difference between the actual pick-up locations and the predictions made by the model on the training data. The MSE for the testing set for pick-ups is approximately 73.93. This indicates the average squared difference between the actual pick-up locations and the predictions made by the model on the testing data. Similarly to drop-offs, a higher MSE suggests less accuracy of the model on the testing data compared to the training data.

Decision Performance:
The average total cost is $28.26. This likely represents the average cost incurred for each ride-hailing trip, including factors such as distance, time, and any additional fees.The average Quality

of Service (QoS) score is 0.88. This suggests that, on average, users rated their ride-hailing experience highly, indicating satisfaction with the service provided. QoS scores are typically based on factors such as driver behavior, vehicle condition, and overall experience during the ride.

Note: For calculating all Decision Performance we have taken number of pickups and drop-offs from the first part of project.

Ridge Regression

The linear model is simpler if we have less features with larger coefficients. Therefore, in our Ridge Regression model, we are using Ridge coefficients as a function of Regularization. Regularization penalizes the large coefficients and improves the conditioning of the problem and reduces the variance of the estimates. Larger values create stronger regularization as shown on the table in our analysis. As the Alpha increases the coefficients all get smaller. We then must use Cross-Validation to find the best Alpha.

Prediction Performance:

The mean squared error (MSE) for the training set for drop-offs is approximately 64.00. This value represents the average squared difference between the actual drop-off locations and the predictions made by the Ridge Regression model on the training data. The MSE for the testing set for drop-offs is approximately 84.39. This value reflects the average squared difference between the actual drop-off locations and the predictions made by the Ridge Regression model on the testing data. A higher MSE for the testing set compared to the training set suggests that the model's performance may degrade when applied to new data. The MSE for the training set for pick-ups is approximately 65.01. This value represents the average squared difference between the actual pick-up locations and the predictions made by the Ridge Regression model on the training data. The MSE for the testing set for pick-ups is approximately 74.39. Similar to drop-offs, this value reflects the average squared difference between the actual pick-up locations and the predictions made by the Ridge Regression model on the testing data.

Decision Performance:

The average total cost is $28.158. This represents the average cost incurred for each ride-hailing trip when using the Ridge Regression model for prediction. Total cost typically includes factors such as distance traveled, time spent, and any additional fees.The average Quality of Service (QoS) score is 0.8825. This score indicates the average level of satisfaction reported by users for their ride-hailing experiences when using the Ridge Regression model. QoS scores are often based on various factors such as driver behavior, vehicle condition, and overall service quality.

Lasso regression:

LASSO (least absolute shrinkage and selection operator)

$\min 12N\sum i=1n(yi-\beta 0-\sum j=1p\beta jxij)2+\alpha\sum j=1p|\beta j|$

· We call the regularization term $L1$ penalty:

$\|\boldsymbol{\beta}\|1=\sum j=1p|\beta j|$

- Zeroes out many coefficients
- Perform an automatic form of feature selection

Lasso regression is a type of linear regression that adds a penalty term to the ordinary least squares objective, which is a sum of the absolute values of the coefficients multiplied by the alpha parameter. This penalty term helps to shrink the coefficients towards zero, effectively performing variable selection by pushing some coefficients to exactly zero.

For pickups:

```
The coefficients are:
temp       -26.893241
precip      -2.611736
uvindex      2.086700
feelslike   30.185890
dtype: float64
```

for drop-offs

```
The coefficients are:
temp       -24.188699
precip      -2.759921
uvindex      2.058665
feelslike   27.614278
dtype: float64
```

Prediction Performance:

The mean squared error (MSE) for the training set for predicting dropoffs is 64.01. This suggests that, on average, the model's predictions for dropoffs deviate from the actual values by approximately 64.01. The MSE for the testing set for predicting dropoffs is 84.38. This indicates that the model's performance slightly worsens when applied to unseen data, as the MSE is higher compared to the training set. The MSE for the training set for predicting pickups is 65.03. Similarly to dropoffs, this suggests that, on average, the model's predictions for pickups deviate from the actual values by approximately 65.03. The MSE for the testing set for predicting pickups is 74.44. Again, similar to dropoffs, the model's performance slightly deteriorates on unseen data compared to the training set.

Decision Performance:

The average total cost is 28.14. This could be a measure of the overall cost associated with the model's predictions. Without further context, it's challenging to interpret this precisely. The quality of service (QoS) score is 0.8828. This indicates the overall performance or accuracy of the model. A score of 1 would represent perfect predictions, so a score of 0.8828 suggests the model is relatively accurate but may still have room for improvement

Overall, these results suggest that the model performs reasonably well in predicting both dropoffs and pickups, with some slight degradation in performance on unseen data. The average total cost and QoS score provide additional insights into the model's overall performance and accuracy.

Elastic NET

- A combination of Ridge Regression and LASSO

$$\min 12N\sum i=1n(yi-\beta 0-\sum j=1p\beta jxij)2+\alpha*\lambda*\sum j=1p|\beta j|+0.5*\alpha*(1-\lambda)\sum j=1p\beta 2j$$

- Here, λ is called 'l1_ratio' in sklearn.

Elastic Net is a regularization technique that combines Lasso and Ridge regression by introducing both L1 and L2 penalties. The alpha value we obtained (0.007544617744578782) indicates the mixing parameter between Lasso and Ridge penalties

For pickups:

The coefficients are:
temp -14.772808
precip -2.664338
uvindex 1.766678
feelslike 18.176608
dtype: float64

For Dropoffs

The coefficients are:
temp -13.070686
precip -2.806305
uvindex 1.766110
feelslike 16.596686
dtype: float64

Prediction Performance:

The mean squared error (MSE) for the training set for predicting dropoffs is 65.60. This suggests that, on average, the model's predictions for dropoffs deviate from the actual values by approximately 65.60. The MSE for the testing set for predicting dropoffs is 87.67. This indicates that the model's performance slightly worsens when applied to unseen data, as the MSE is higher compared to the training set. The MSE for the training set for predicting pickups is 66.87. Similarly to dropoffs, this suggests that, on average, the model's predictions for pickups deviate from the actual values by approximately 66.87. The MSE for the testing set for predicting pickups is 77.61. Again, similar to dropoffs, the model's performance slightly deteriorates on unseen data compared to the training set.

Decision Performance:

The average total cost is 31.05. This could be a measure of the overall cost associated with the model's predictions. Without further context, it's challenging to interpret this precisely. The average quality of service (QoS) score is 0.930. This indicates the overall performance or accuracy of the model. A score of 1 would represent perfect predictions, so a score of 0.930 suggests the model is relatively accurate but may still have room for improvement.

Overall, these results suggest that the Elastic NET model performs reasonably well in predicting both dropoffs and pickups, with some slight degradation in perform. ance on unseen data. The average total cost and QoS score provide additional insights into the model's overall performance and accuracy.

KNN

In employing the K-nearest neighbors (KNN) algorithm for predictive modeling, we assess its performance using key metrics. KNN is a simple yet effective algorithm that classifies objects based on their similarity to neighboring data points. Here's an interpretation of the results obtained from applying KNN to predict dropoffs and pickups:

Prediction Performance:
Mean Squared Error (MSE) for Testing Set: The MSE for dropoffs on the testing set is 108.89, indicating that, on average, the model's predictions for dropoffs deviate from the actual values by approximately 108.89 units. Similarly, for pickups, the MSE on the testing set is 88.57, suggesting a slightly lower deviation in predictions compared to dropoffs.

Decision Performance:
Average Total Cost: The average total cost associated with the KNN model's predictions is 35.58. This metric provides insight into the economic implications of the model's performance, although further context is required to interpret it fully.

Average Quality of Service (QoS) Score: The average QoS score for the KNN model is 0.958, suggesting a high level of accuracy in its predictions. A score of 1 represents perfect predictions, so a score of 0.958 indicates that the model is highly accurate, with minimal room for improvement.

Overall, the KNN model demonstrates strong predictive performance, as evidenced by its low MSE values and high QoS score. However, further analysis may be needed to understand the implications of its predictions on practical applications, particularly concerning the average total

---

Overall Summary:

Methodology
1)Here We have found the MSE of training and Test for dropoffs and pickups.
2)Then predicticted the values.
3)Then Found Total cost for each test point then have taken avg of those total cost.
4)Then Found AVG QOS for each test point then have taken avg of those QOS.

Combining all Model Performance

Linear regression
For Linear Regression
1.MSE for training set for dropoffs: 63.967535226608724
2.MSE for testing set for dropoffs: 84.14926073268506
3.MSE for training set for pickups: 64.98246773898656
4.MSE for testing set for pickups: 73.93266405254103
5.The Avg total cost is 28.26
6.The Avg Qos Score is 0.88

Ridge regression
For Ridge Regression
1.MSE for training set for dropoffs: 64.00317424004356
2.MSE for testing set for dropoffs: 84.3946056705549
3.MSE for training set for pickups: 65.01391582914607
4.MSE for testing set for pickups: 74.38849145908942
5.The Avg total cost is 28.158
6.The Avg Qos Score is 0.8825

LASSO
For LASSO
1.MSE for training set for dropoffs: 64.01242167578064
2.MSE for testing set for dropoffs: 84.37626472621518
3.MSE for training set for pickups: 65.02734234287406
4.MSE for testing set for pickups: 74.4415462406801
5.The Avg total cost is 28.1375
6.The Avg Qos Score is 0.8828

Elastic NET
For Elastic NET
1.MSE for training set for dropoffs: 65.59719110525381
2.MSE for testing set for dropoffs: 87.67008431171384
3.MSE for training set for pickups: 66.8661335837862
4.MSE for testing set for pickups: 77.60870874551749
5.The Avg total cost is 31.0468
6.The Avg Qos Score is 0.930

KNN
For KNN
1.MSE for testing set for dropoffs: 108.8875
2.MSE for testing set for pickups: 88.5745
3.The Avg total cost is 35.58
4.The Avg Qos Score is 0.958

Conclusion & Limitations:

Based on our results the LASSO regression model is good choice because it's both simple and performs reasonably well. But even though it seems good, the testing we did didn't show super accurate results for all models. This suggests that if we add more stuff, like information about time and where things are happening, we might be able to make better predictions. Another thing is, we only looked at one spot where people get picked up and dropped off, at 22nd & H St NW. It's important to look at more places to make sure our findings are solid.
In the future, we think it's a good idea to try out fancier models and include more details about what's going on. We could also try adjusting certain settings to see if that helps. Plus, if we bring

in extra data from other sources, we might learn even more. Our study is just the start of figuring out how people use ride-hailing services, and we hope it leads to better ways of predicting things.

Author Contributions

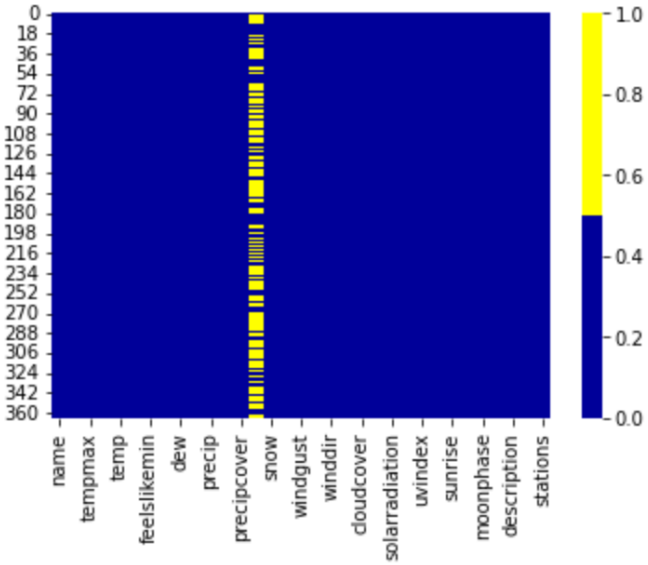
Part1.Initial discussions about formulating problem and modeling approaches: All
Part2.Model training and evaluation: Sahith
Part3.Analysis: Nour Alzaid
Part4.Compiled and wrote the final report : Shubham Patil, Kason Richard

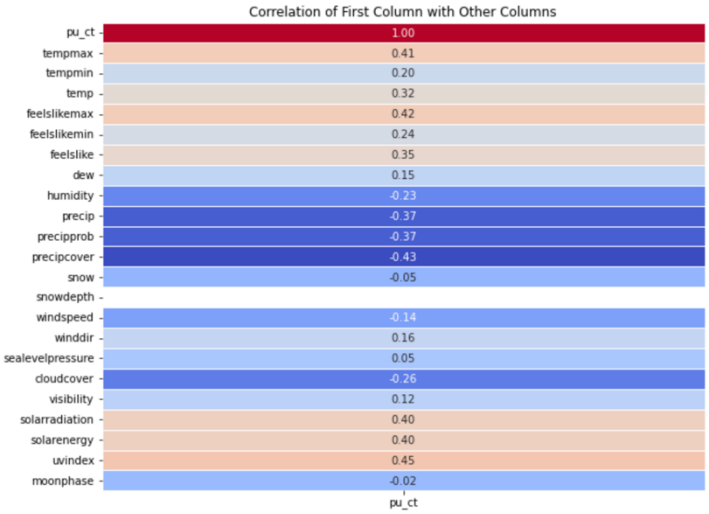
---

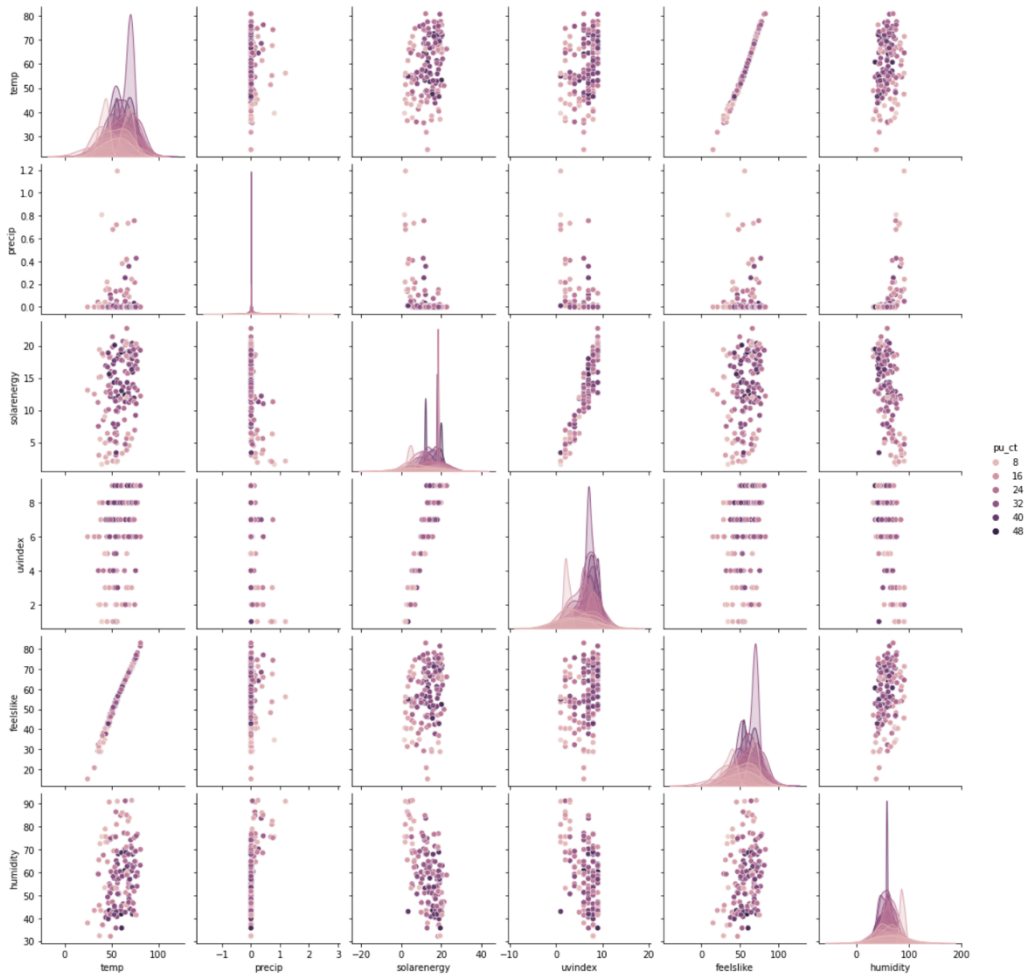
You can access our whole code file here
https://drive.google.com/file/d/1lnuVw4THcmbLWmRuZZ2wuqC6k3BeQF_H/view?usp=share_link

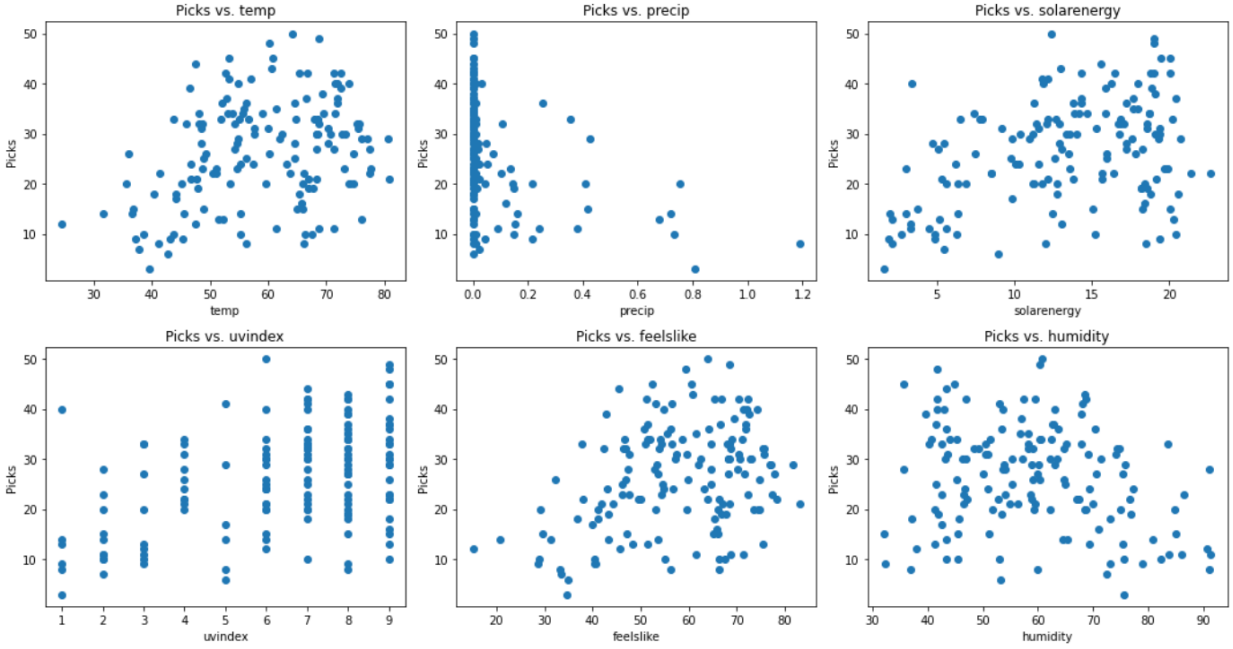
---

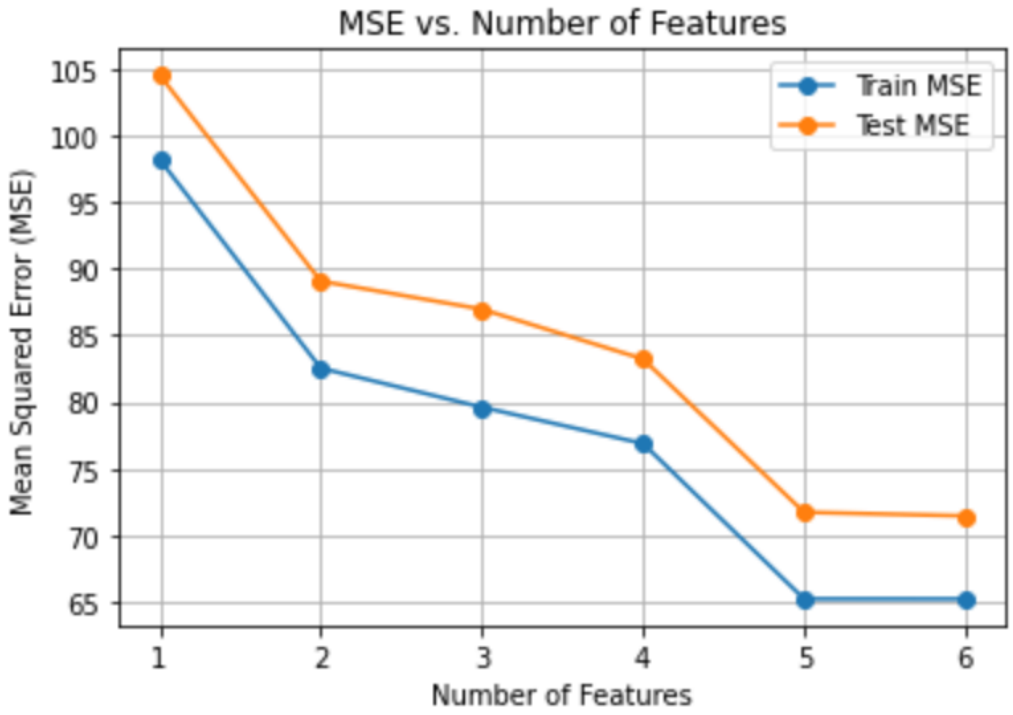
Appendix

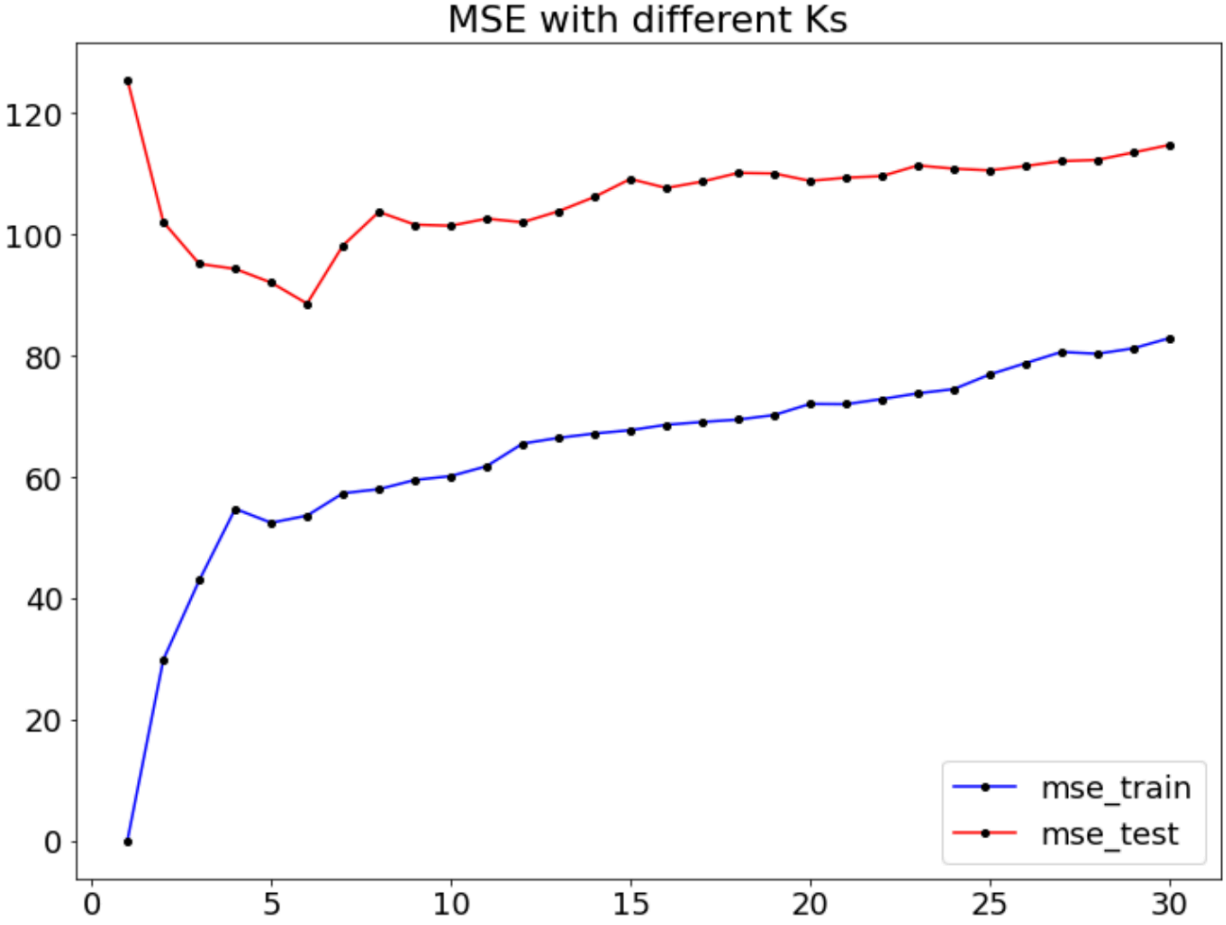


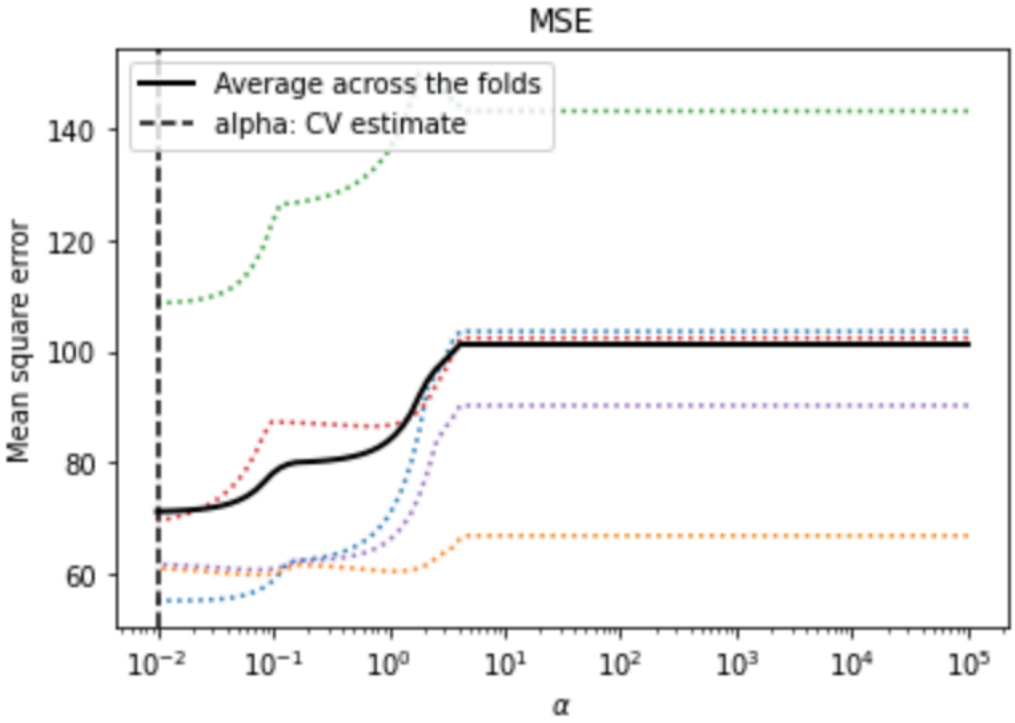


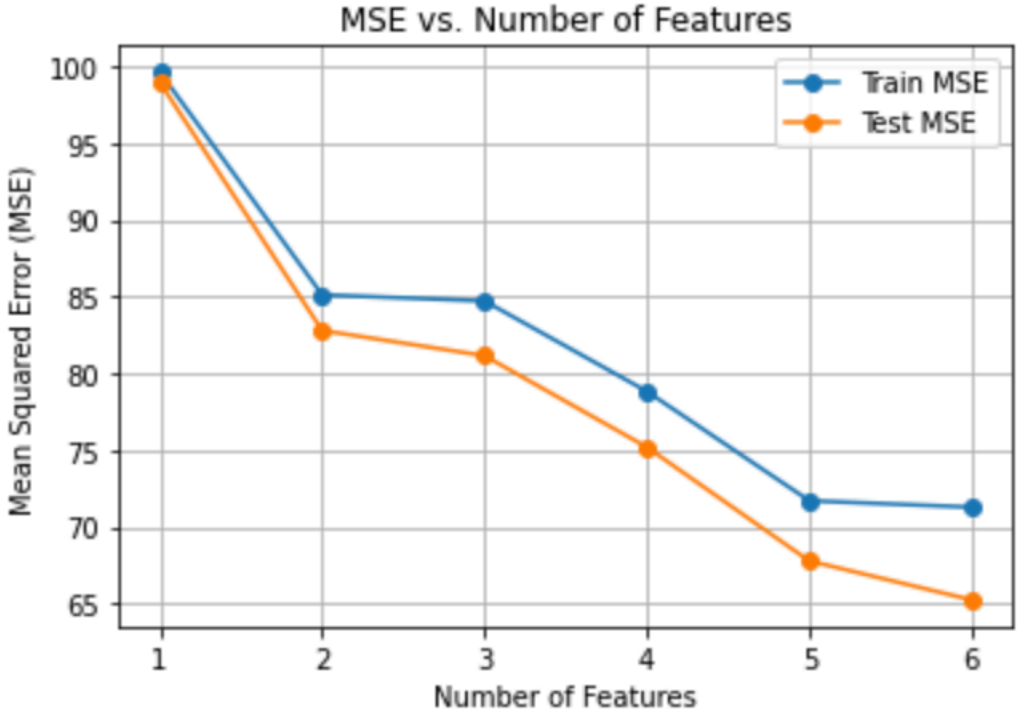

temp
precip
solarenergy
uvindex
feelslike
humidity
pu_ct
8
16
24
32
40
48

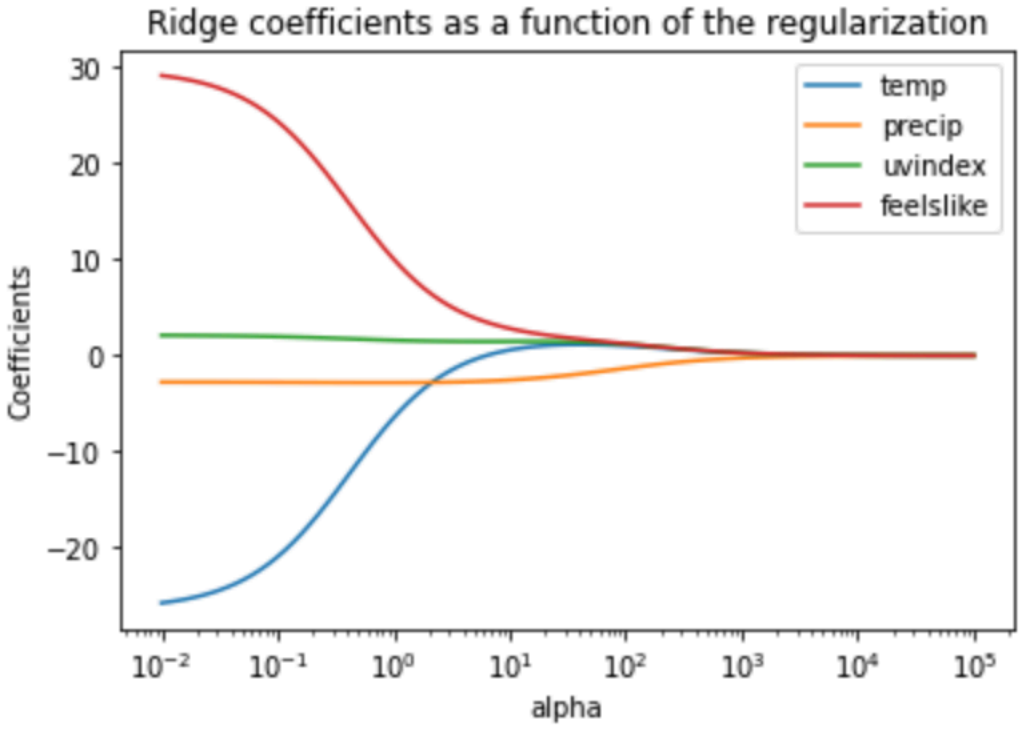

Picks vs. temp
Picks vs. precip
Picks vs. solarenergy
Picks vs. uvindex
Picks vs. feelslike
Picks vs. humidity
Picks
temp
precip
solarenergy
uvindex
feelslike
humidity


MSE vs. Number of Features
Train MSE
Test MSE
Mean Squared Error (MSE)
Number of Features
105
100
95
90
85
80
75
70
65
1
2
3
4
5
6


MSE with different Ks
mse_train
mse_test
120
100
80
60
40
20
0
0
5
10
15
20
25
30


MSE
Average across the folds
alpha: CV estimate
Mean square error
140
120
100
80
60
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
α


MSE vs. Number of Features
Train MSE
Test MSE
Mean Squared Error (MSE)
100
95
90
85
80
75
70
65
1
2
3
4
5
6
Number of Features


Ridge coefficients as a function of the regularization
Coefficients
alpha
temp
precip
uvindex
feelslike
30
20
10
0
−10
−20
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}