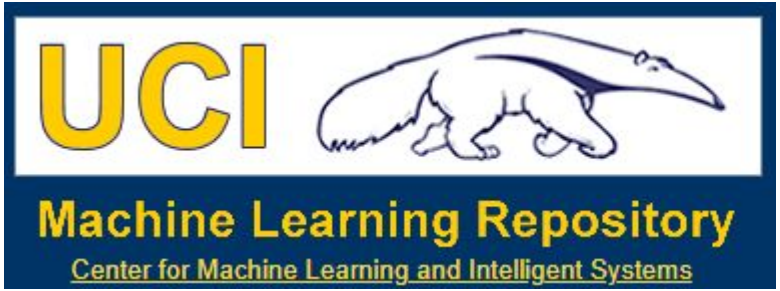
DS 5220 Final Project

Shivani Patel and Owen Davey

The Data Set


UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

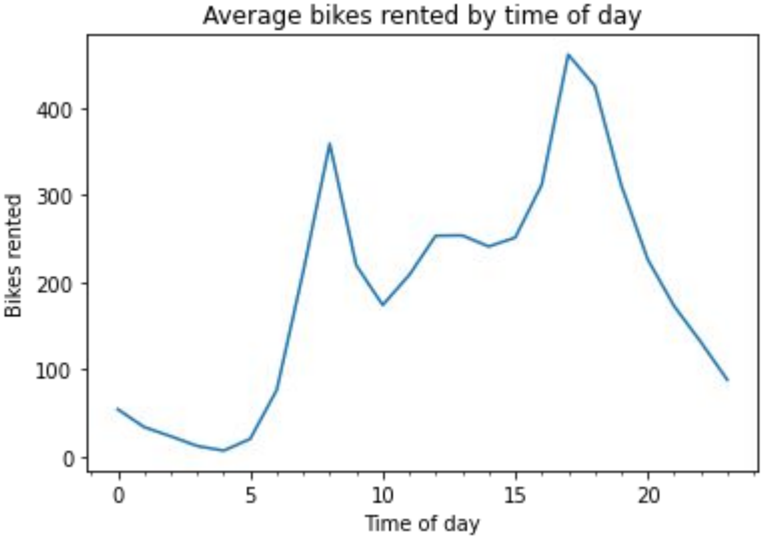
- The Bike Sharing Data Set from the UCI Machine Learning Repository
  - Measures the number of bike rentals from a automated bike rental stand in a given hour
  - 17389 instances
  - 16 Atributes

- Attributes
  - Record Index
  - Date
  - Season (1:winter, 2:spring, 3:Summer, 4:Fall)
  - Year
  - Month (1 to 12)
  - Hour (1 to 23)
  - Holiday
  - Weekend
  - Working Day
  - Weather
  - Temperature
  - Humidity
  - Windspeed
  - Casual rentals
  - Register user rentals
  - Count of total rentals

Data Cleaning

1. Trimmed off Index and membership information
2. Created a Boolean attribute to represent high rentals
3. Split X and Y

Analysis

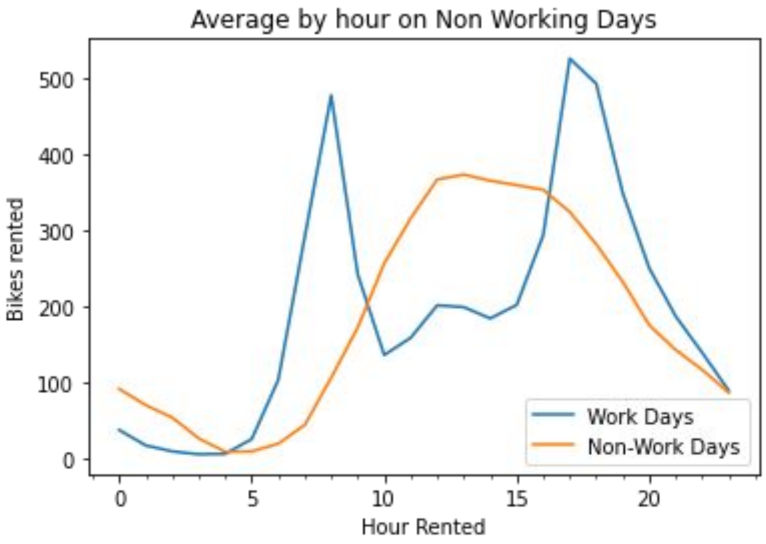


We performed some rudimentary analysis on the data.

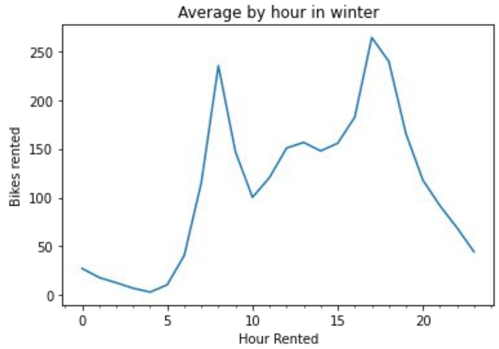
Here we see that during peak commuting hours our bike rentals also peak.

Non working days

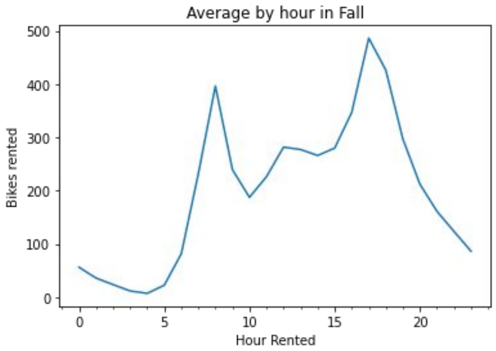
This is a study of how rentals change when it is a standard working day (Monday-Friday) vs a standard non-working day (Weekends and Holidays).

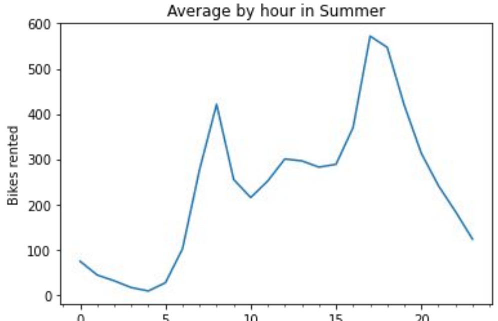


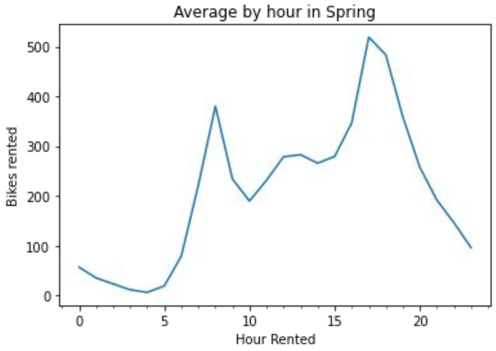
When we control for season we see



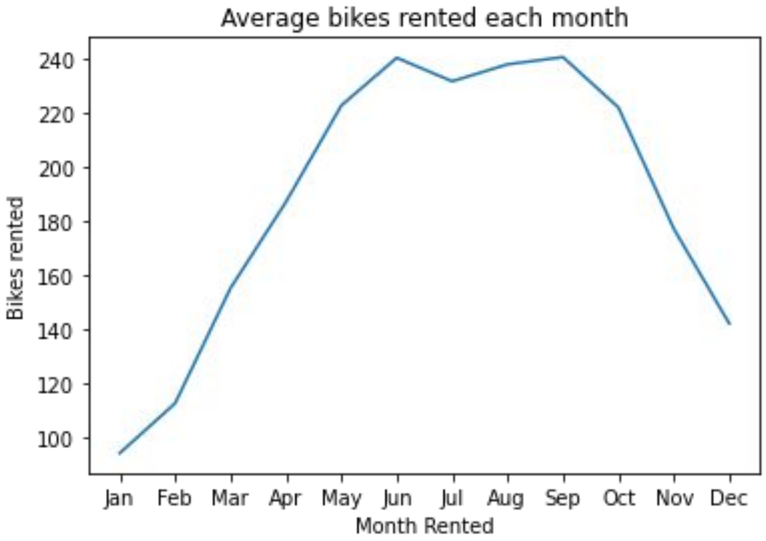
Though the actual amount of rentals at a given time fluctuates the peak hours remain the same with extremely similar shapes.





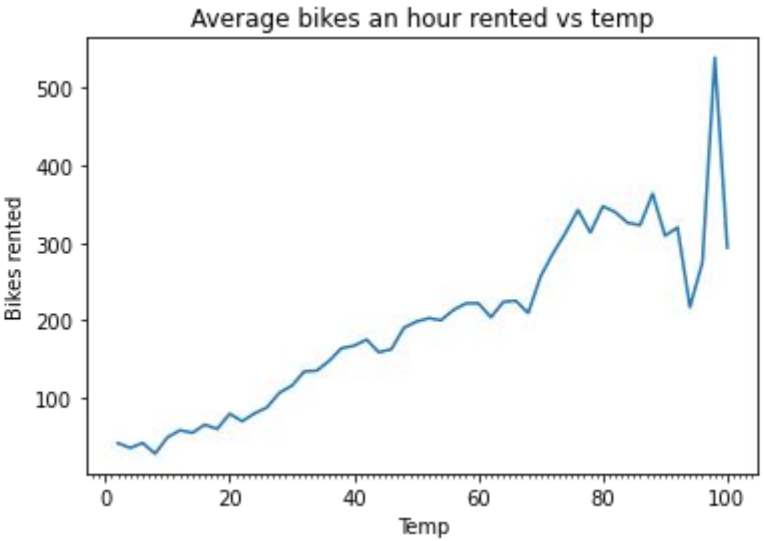


Bikes rented by month



We can see there does seem to be a strong correlation between month and rentals with the highest average hourly rentals happening in the summer months

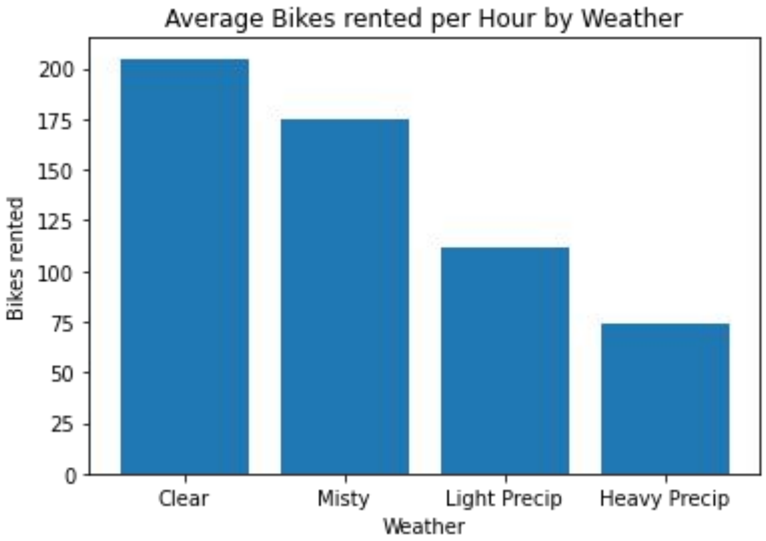
How temperature effects



Given the relationship between months and rentals we specifically explored temperature and bike rentals. The instability in the graph towards the end can be explained by having fewer data points.

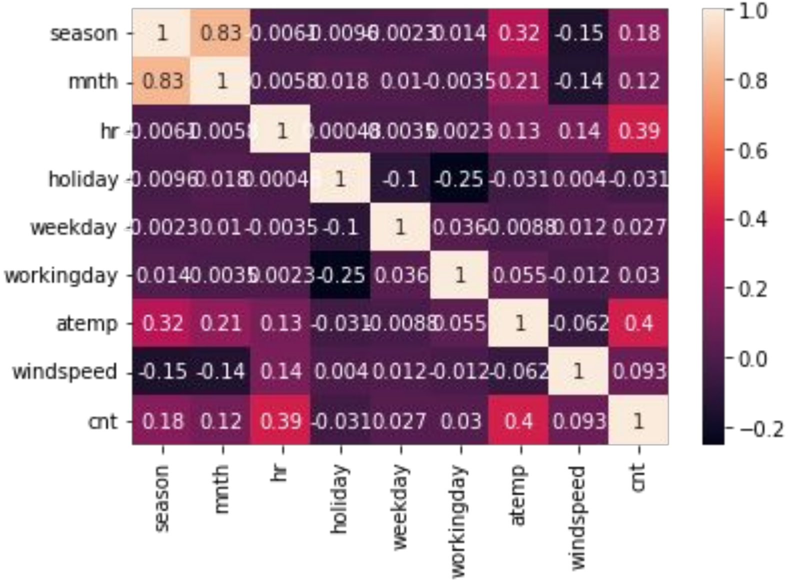
Weather

The next step was to look at how weather affects the number of bikes rented. We can see and decrease in rentals as weather gets worse



Correlation Matrix

Finally, we chose to do a correlation matrix on our data. There isn't a strong correlation between any of the variables that isn't obvious (season and month), which means that when predicting the number of bike rentals, looking at any individual feature will not be sufficient

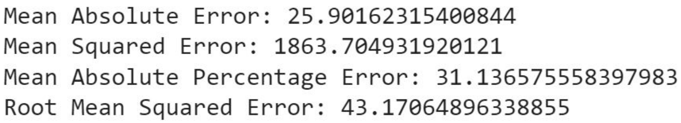


Methodology

- Random Forest Regression
- Random Forest Classification
- Naive Bayes
- Neural Network
- Logistic Regression
- Ridge Regression

Random Forest (Regression & Classification)

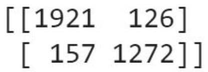
Regression:



```
Mean Absolute Error: 25.90162315400844
Mean Squared Error: 1863.704931920121
Mean Absolute Percentage Error: 31.136575558397983
Root Mean Squared Error: 43.17064896338855
```

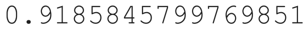
Classification:

Confusion Matrix:



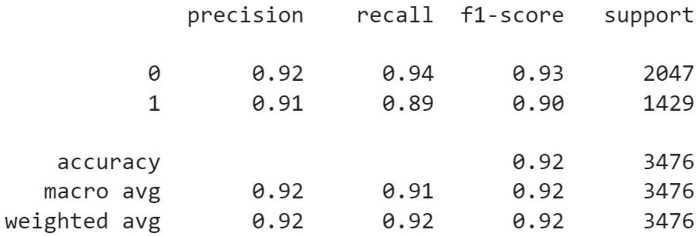
```
[[1921  126]
 [ 157 1272]]
```

Accuracy Score:



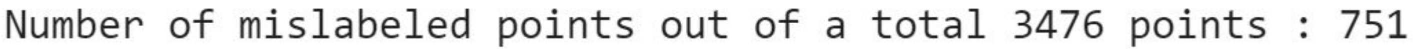
```
0.9185845799769851
```

Classification Report:



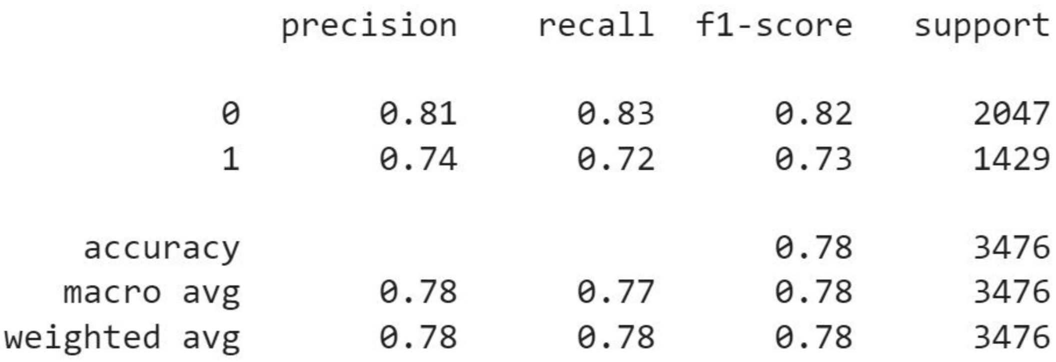
```
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      2047
           1       0.91      0.89      0.90      1429

    accuracy                           0.92      3476
   macro avg       0.92      0.91      0.92      3476
weighted avg       0.92      0.92      0.92      3476
```

Naive Bayes



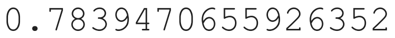
```
Number of mislabeled points out of a total 3476 points : 751
```

Classification Report:



```
              precision    recall  f1-score   support

           0       0.81      0.83      0.82      2047
           1       0.74      0.72      0.73      1429

    accuracy                           0.78      3476
   macro avg       0.78      0.77      0.78      3476
weighted avg       0.78      0.78      0.78      3476
```

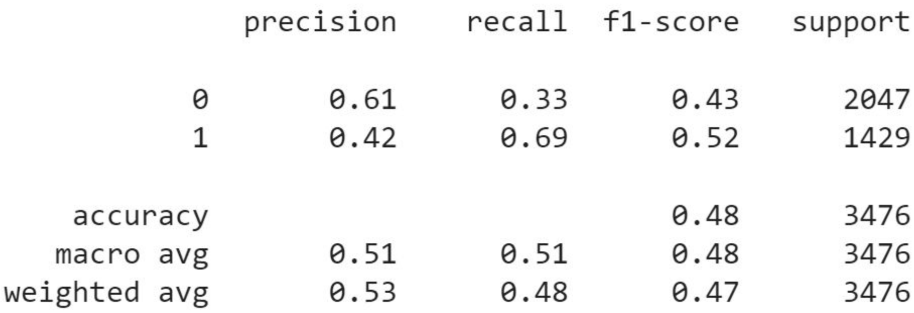
Accuracy Score:



```
0.7839470655926352
```

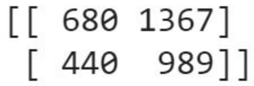
Neural Networks

Classification Report:



```
              precision    recall  f1-score   support

           0       0.61      0.33      0.43      2047
           1       0.42      0.69      0.52      1429

    accuracy                           0.48      3476
   macro avg       0.51      0.51      0.48      3476
weighted avg       0.53      0.48      0.47      3476
```

Confusion Matrix:



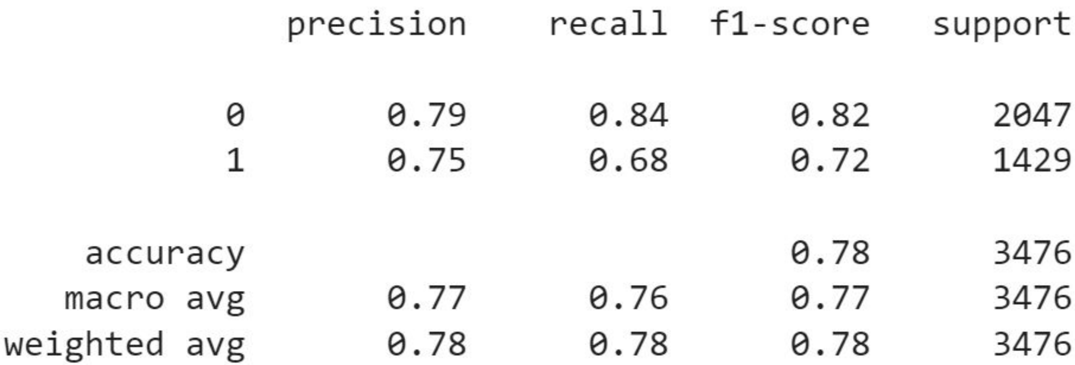
```
[[ 680 1367]
 [ 440  989]]
```

Accuracy Score:

```
0.48014959723820483
```

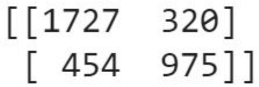
Log Regression

Classification Report:



```
              precision    recall  f1-score   support

           0       0.79      0.84      0.82      2047
           1       0.75      0.68      0.72      1429

    accuracy                           0.78      3476
   macro avg       0.77      0.76      0.77      3476
weighted avg       0.78      0.78      0.78      3476
```

Confusion Matrix:



```
[[1727  320]
 [ 454  975]]
```

Accuracy Score:

```
0.7773302646720368
```

Ridge Regression

Training Data:

RMSE:

```
141.47307170557784
```

R2 Score:

```
0.38730099774858595
```

Testing Data:

RMSE:

```
143.1306938002819
```

R2 Score:

```
0.3944740239656034
```

Results

Of all of the tried methods, random forest classification had the highest accuracy and precision scores and ridge regression had an r2 score that was lower than 0.4, making it a bad model for our data.

Future Steps:
Implement K-Fold and cross validation

Take a closer look at registered vs unregistered bike rentals