# Predicting trends in bike share program usage

Team: Zixuan Wang, Shuqiong Chen, Kevin Walsh
Instructor: Amir H. Gandomi

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

## Introduction

**Problem:**

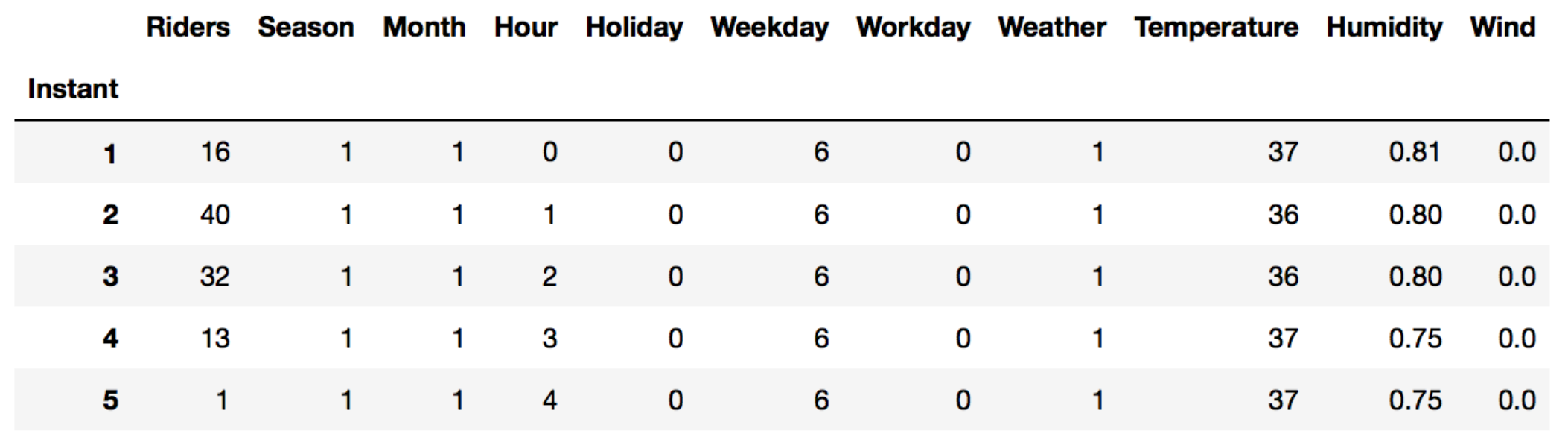
- Predict the amount of riders under different condition by multiple linear regression and multiple polynomial regression
- Predict whether the bikes would be heavily used on certain times

**Business value:**

- This model will help companies to distribute theirs bikes in an optimal way.
- Companies will reduce costs incurred through bike dispatch and help increase their business brand awareness and retention rate of customers.

## Data understanding and processing

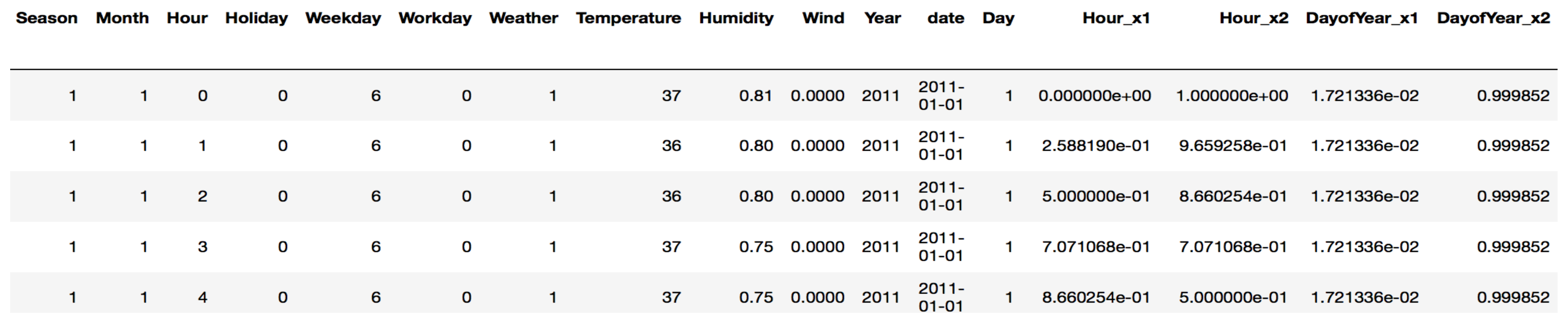
- Dataset: 17,379 bike share record entries with 11 variables, including continuous, binary, and categorical variables.



| Instant | Riders | Season | Month | Hour | Holiday | Weekday | Workday | Weather | Temperature | Humidity | Wind |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 16 | 1 | 1 | 0 | 0 | 6 | 0 | 1 | 37 | 0.81 | 0.0 |
| 2 | 40 | 1 | 1 | 1 | 0 | 6 | 0 | 1 | 36 | 0.80 | 0.0 |
| 3 | 32 | 1 | 1 | 2 | 0 | 6 | 0 | 1 | 36 | 0.80 | 0.0 |
| 4 | 13 | 1 | 1 | 3 | 0 | 6 | 0 | 1 | 37 | 0.75 | 0.0 |
| 5 | 1 | 1 | 1 | 4 | 0 | 6 | 0 | 1 | 37 | 0.75 | 0.0 |

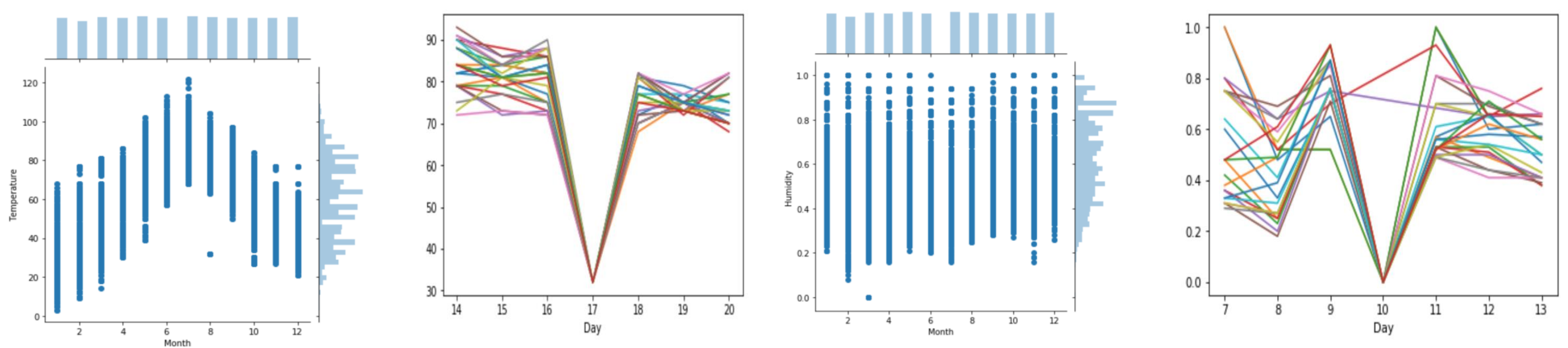
- Data insights: The whole dataset contains records from 724 continuous days. From the date information of this dataset, we found that this dataset contains records from 2011-2012.

- Feature engineering: Understanding the dataset helps us extract more information from the raw data. Now we can impute date and year label for each row. For periodic features, such as hour and day, we can use polar coordinates to transform them, so each point can be calculated through trigonometric functions. In this way, we can replace the "Month" and "Season" variables.



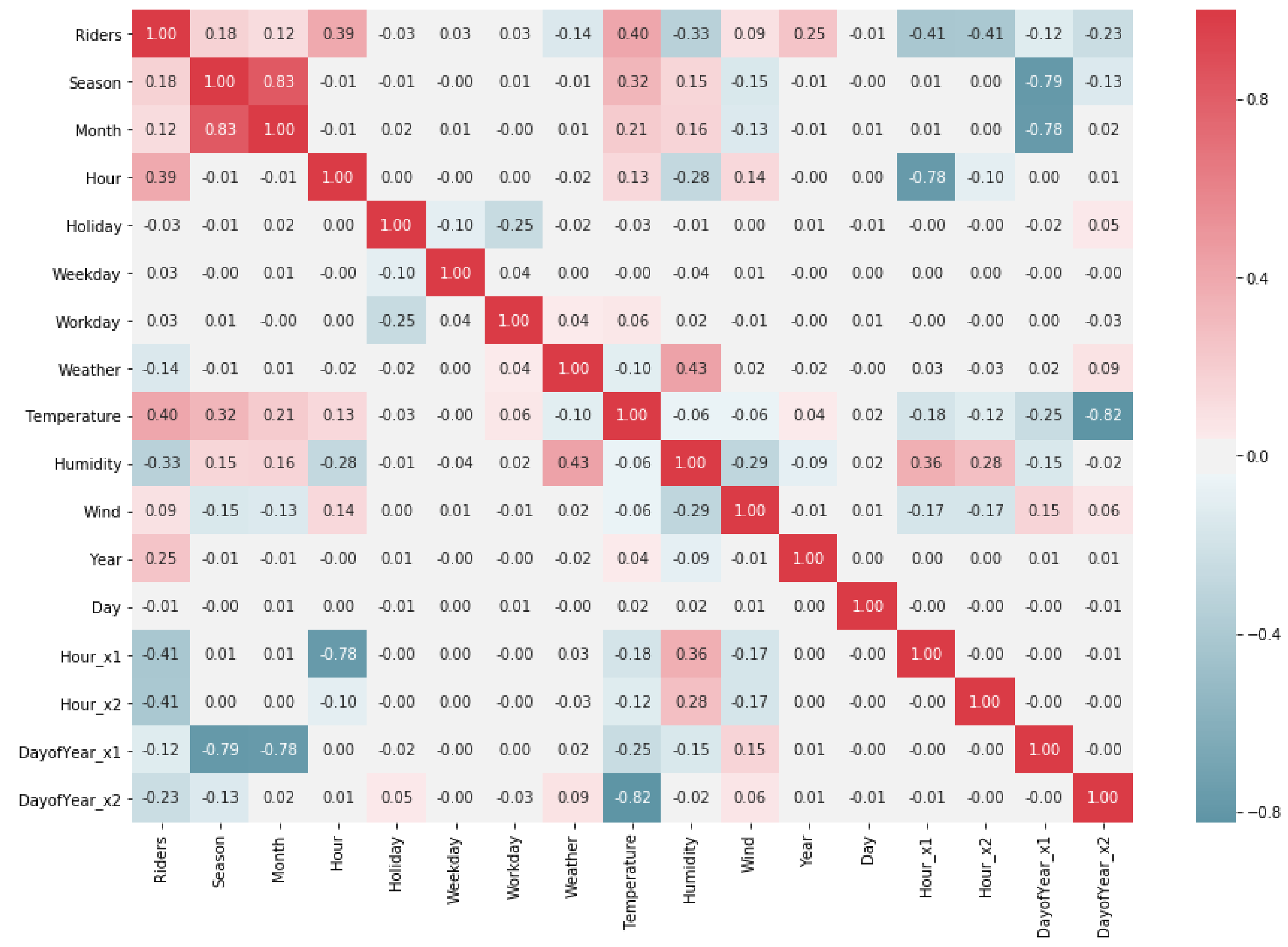
| Season | Month | Hour | Holiday | Weekday | Workday | Weather | Temperature | Humidity | Wind | Year | date | Day | Hour_x1 | Hour_x2 | DayofYear_x1 | DayofYear_x2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 6 | 0 | 1 | 37 | 0.81 | 0.0000 | 2011 | 2011-01-01 | 1 | 0.000000e+00 | 1.000000e+00 | 1.721336e-02 | 0.999852 |
| 1 | 1 | 1 | 0 | 6 | 0 | 1 | 36 | 0.80 | 0.0000 | 2011 | 2011-01-01 | 1 | 2.588190e-01 | 9.659258e-01 | 1.721336e-02 | 0.999852 |
| 1 | 1 | 2 | 0 | 6 | 0 | 1 | 36 | 0.80 | 0.0000 | 2011 | 2011-01-01 | 1 | 5.000000e-01 | 8.660254e-01 | 1.721336e-02 | 0.999852 |
| 1 | 1 | 3 | 0 | 6 | 0 | 1 | 37 | 0.75 | 0.0000 | 2011 | 2011-01-01 | 1 | 7.071068e-01 | 7.071068e-01 | 1.721336e-02 | 0.999852 |
| 1 | 1 | 4 | 0 | 6 | 0 | 1 | 37 | 0.75 | 0.0000 | 2011 | 2011-01-01 | 1 | 8.660254e-01 | 5.000000e-01 | 1.721336e-02 | 0.999852 |

- Outliers:

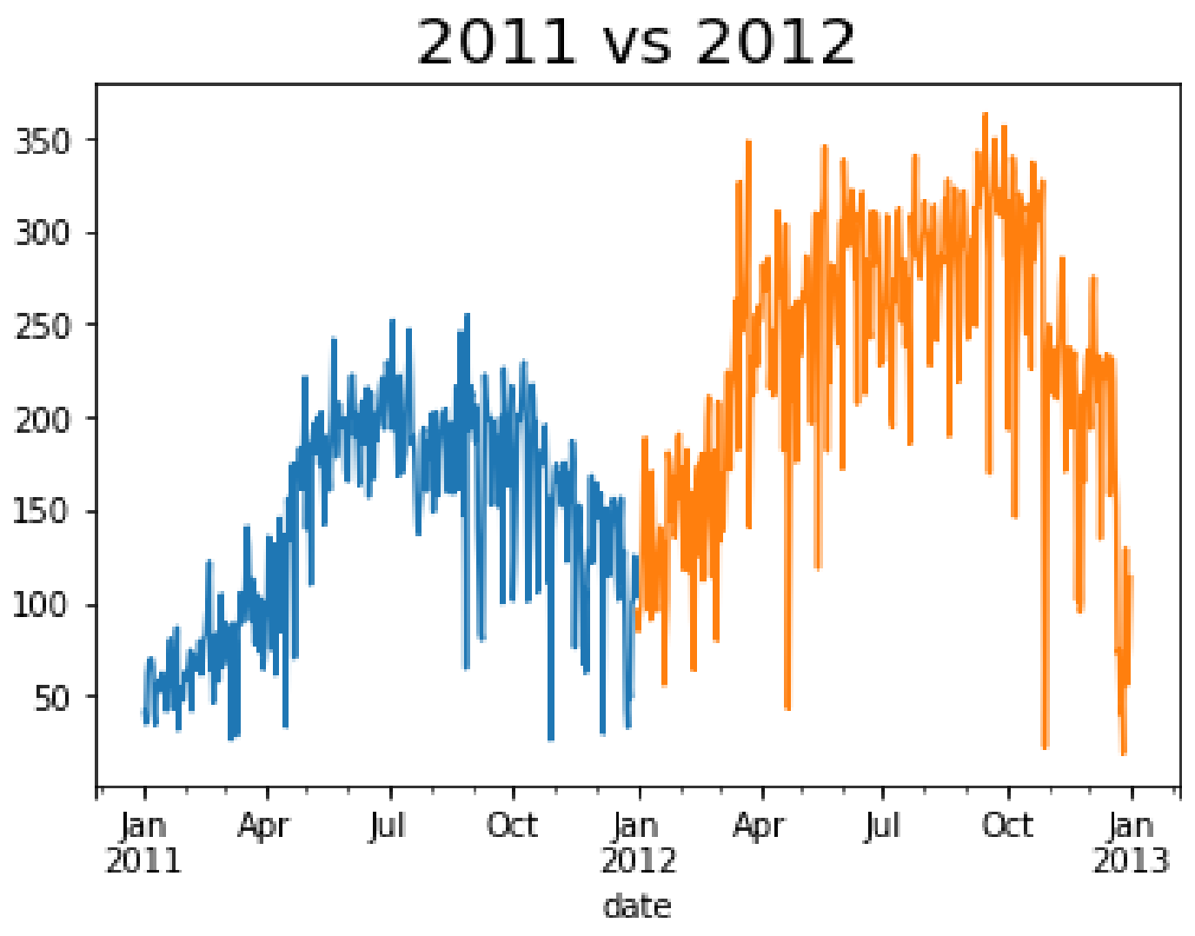


- From plots comparing different variables, we found several outliers. After investigating detailed data around outlier records, we used various methods to adjust these outliers.

•Correlation Coefficients



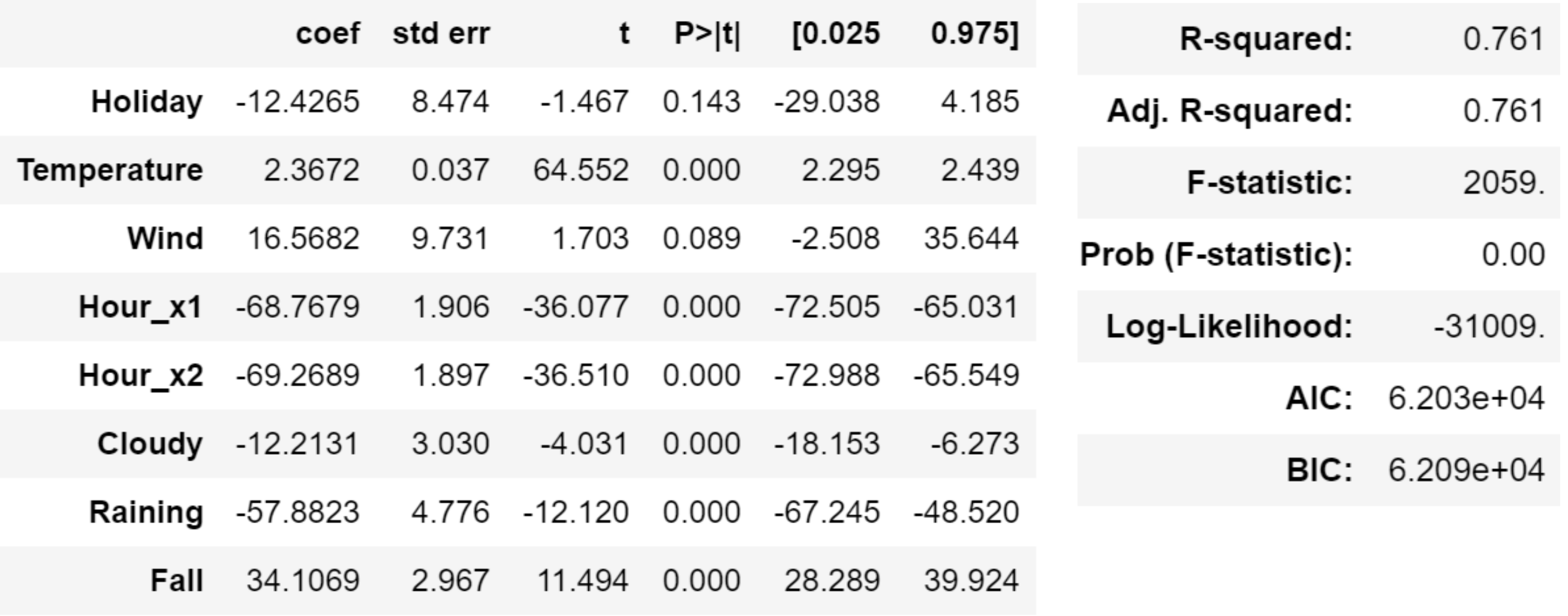
## Modelling



Regression: we found a linear connection between 2011 and 2012 data. So, we will build a model using 2011 data to predict the number of riders in 2012.

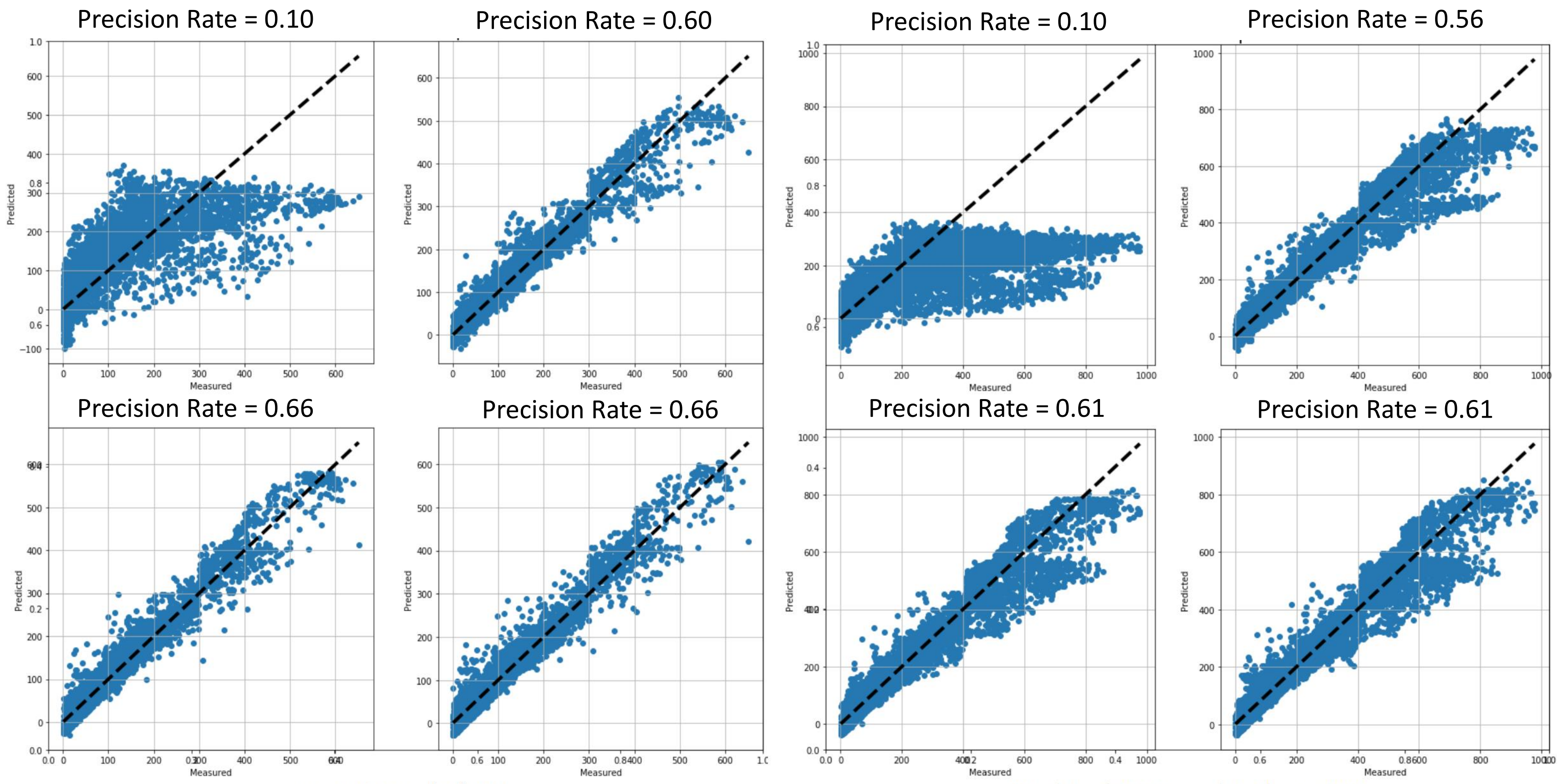
Classification: we aim to classify busy or idle usage. We define busy use as greater than 243 riders in 2011 and 400 riders in 2012.

- Regression:



| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Holiday | -12.4265 | 8.474 | -1.467 | 0.143 | -29.038 | 4.185 |
| Temperature | 2.3672 | 0.037 | 64.552 | 0.000 | 2.295 | 2.439 |
| Wind | 16.5682 | 9.731 | 1.703 | 0.089 | -2.508 | 35.644 |
| Hour_x1 | -68.7679 | 1.906 | -36.077 | 0.000 | -72.505 | -65.031 |
| Hour_x2 | -69.2689 | 1.897 | -36.510 | 0.000 | -72.988 | -65.549 |
| Cloudy | -12.2131 | 3.030 | -4.031 | 0.000 | -18.153 | -6.273 |
| Raining | -57.8823 | 4.776 | -12.120 | 0.000 | -67.245 | -48.520 |
| Fall | 34.1069 | 2.967 | 11.494 | 0.000 | 28.289 | 39.924 |

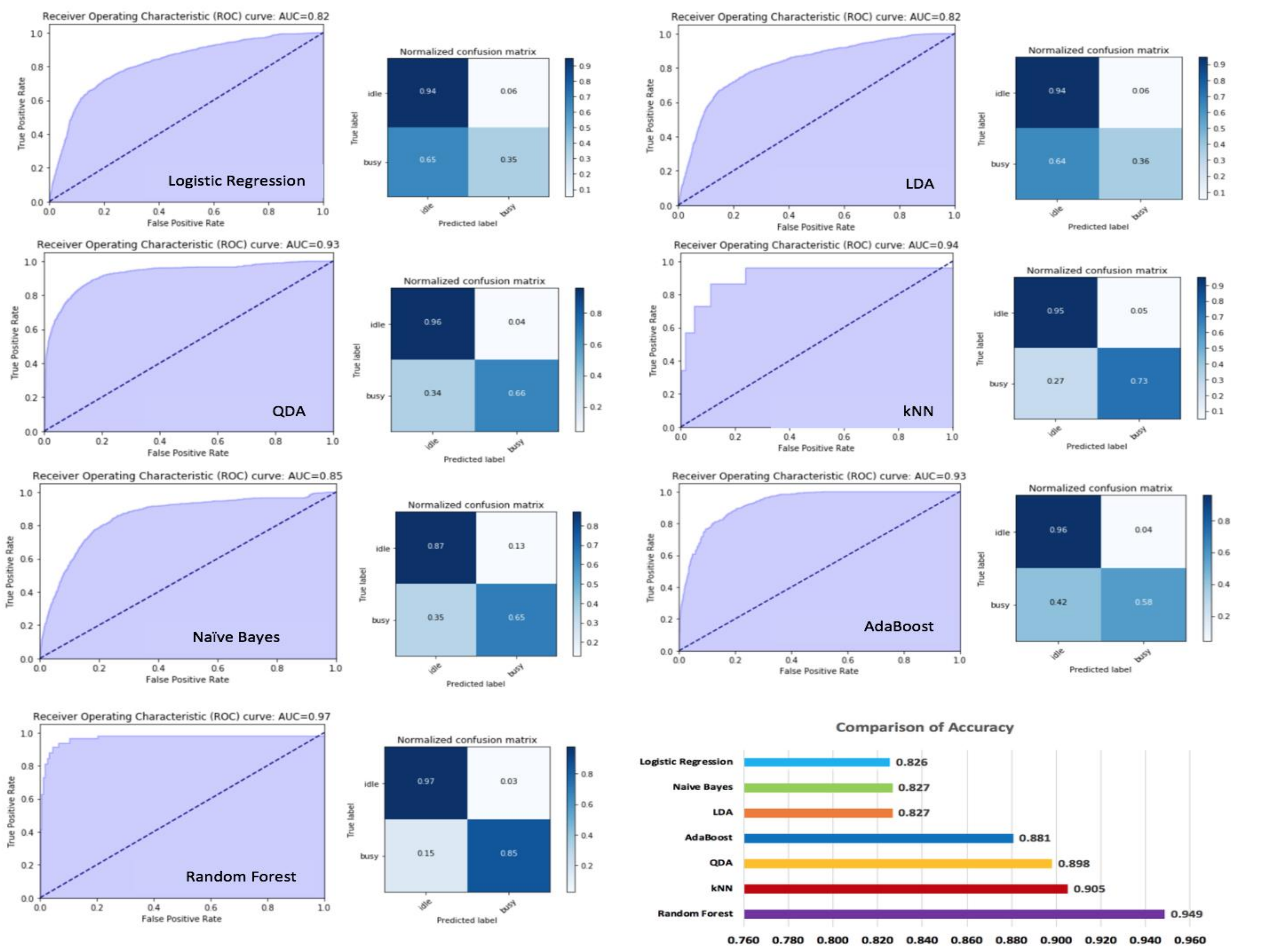
| | |
|---|---|
| R-squared: | 0.761 |
| Adj. R-squared: | 0.761 |
| F-statistic: | 2059. |
| Prob (F-statistic): | 0.00 |
| Log-Likelihood: | -31009. |
| AIC: | 6.203e+04 |
| BIC: | 6.209e+04 |



Regression result of 2011

Use data of 2011 to predict riders in 2012

- Classification:



## Conclusion & Future Work

- Our regression model is highly descriptive. People tend to prefer using bikes in warmer weather during daytime.
- Random Forest provides the best result for classification problem.
- The number of riders in 2012 was noticeably increased compared to the number of riders in 2011. We suggest that the company analyze differences in their operational strategy, since there is little difference in external factors between 2011 and 2012.