# San Franscisco bike share rentals
# Using Microsoft Azure and Databricks

ANUPAM SAHAY
KRITHY NANAIAH ATRANGADA
NEHA SHASHIDHARA GULI
Department of Information Systems, California State University
Los Angeles
e-mail : asahay@calstatela.edu, katrang@calstatela.edu, sneha@caltstatela.edu

**Abstract:** This project will illustrate the usage of Microsoft Azure and Databricks on San Franscisco bike share rentals dataset.we will utilize the knowledge learnt in class, extensive researches and development of predictive model uisng Microsoft Azure in order to predict the number of trips taken on a particular day. In Databricks we have used the Spark ML to find the model accuracy and root mean square error. We have used three algorithms to compare the accuracy in databricks.

URL:https://www.kaggle.com/benhamner/sf-bay-area-bike-share/data
Dataset Size:4.45GB

Data Bricks Specifications
Execution: Single Node
Memory: 6GB Capacity
Databricks Runtime Version: 4.0(Incl. Apache Spark 2.3.0, Scala 2.11)

Microsoft Azure Machine Learning Studio Specification
Execution: Single Node
Storage Space Capacity: 10 GB Memory
Compute Resource Type: ML service is a multitenant service.

## 1. Introduction

Based on generous list of data provided by our instructor, we have done some researches and excusively decided which data we are using for this project.we are going to use the dataset to do prediction.

- We have created a model to predict the number of trips taken on a particular day.
- We are using all the information and data provided by the SF bike sharing company i.e. weather report, number of bikes available, type of days(Business day v/s Holiday v/s Weekends )
- We treat Business days as 1 and Holidays or Weekends as 0
- We have checked if the stations were repeated or listed only once
- We have split the data in train and test in the ratio of 0.8 and 0.2
- We have checked for distinct weather events i.e. rain, fog, normal day etc.
- We have checked for any outliers in the data.

## 2. Predective Modelling

The spark Machine Learning has been used in Databricks to find out the model accuracy with cross validation and to find the root mean Square error. Three different algorithms have been used to compare the models accurecy and to find the best fit model.
The same model has been built for prediction in Microsoft Azure.
The following sections provides the model and the result.

### 2.1 DATA Bricks (SPARK Machine Learning)

**MODEL 1:**
ALGORITHM USED : **DECISION TREE REGRESSION**
MODEL ACCURACY WITH CROSS VALIDATION : **90.295**
ROOT MEAN SQUARE ERROR: **147.739**

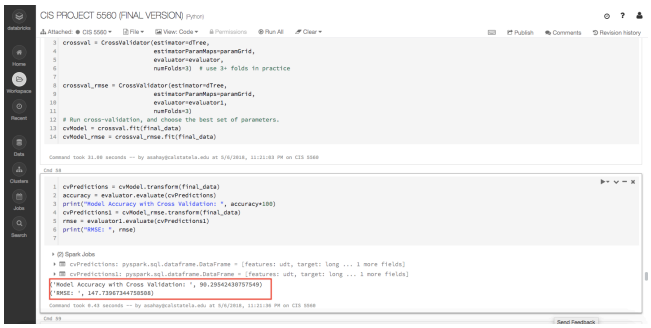


**Figure 1. Decision Tree Regression**

**MODEL 2:**

ALGORITHM USED: **RANDOM FOREST REGRESSION**

MODEL ACCURACY WITH CROSS VALIDATION: **94.8094**

ROOT MEAN SQUARE ERROR: **92.12**

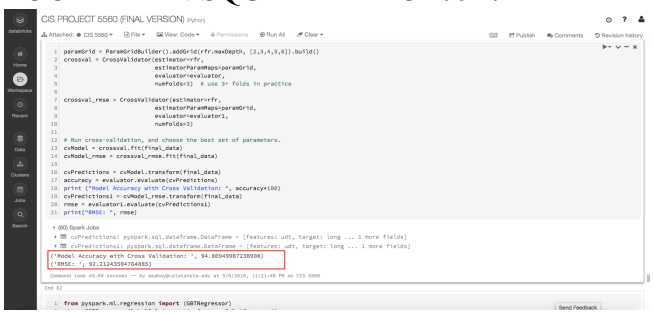


**Figure 2. Random Forest Regression**

**MODEL 3:**

ALGORITHM USED: **GRADIENT BOOSTING REGRESSION**

MODEL ACCURACY WITH CROSS VALIDATION: **94.3279**

ROOT MEAN SQUARE ERROR: **96.3950**

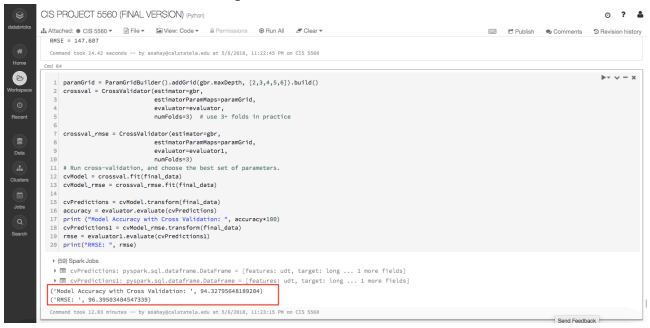


**Figure 3. Gradient Boosting Regression**

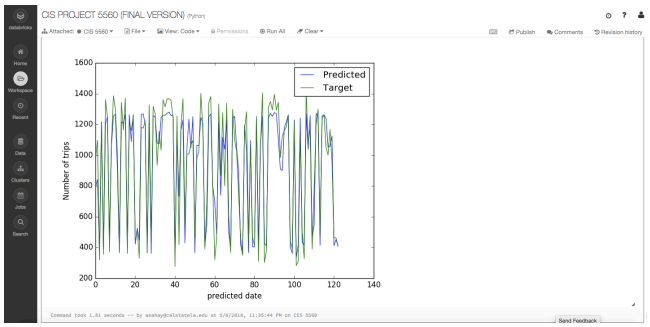


**Figure 4. Graph: Random forest regression graph**

- This graph predicts the the target trips per day v/s the predicted number of trips.
- This predictions should give the bike sharing company about the usage of bike trips on a particular day depending on weather forecast, type of day(Business day or Holiday), number of bikes available on a particular station etc.
- It will give the company an estimate of traffic congestion that will occur each day.
- Future work of the project is that we can create a similar model or a new model to predict the number of trips from each station, predicting when the station will run out of bikes, what will be status of bike in any particular day.
- We can also forecast number of trips in the morning to predict number of trips in the afternoon.

## 2.2 MICROSOFT AZURE MACHINE LEARNING STUDIO

Here we have used two algorithms to build the model and predict the outcome.

One is Boosted decision tree regression and the other decision forest regression

The data source input has been taken from three sources: trip.csv, weather.csv and station.csv.

Execute python script modle was used to sample the data.

Next select columns in dataset model was used to select the columns which will be used in prediction. Later data cleaning was done and an algorithm was applied to verify the model for its accuracy.

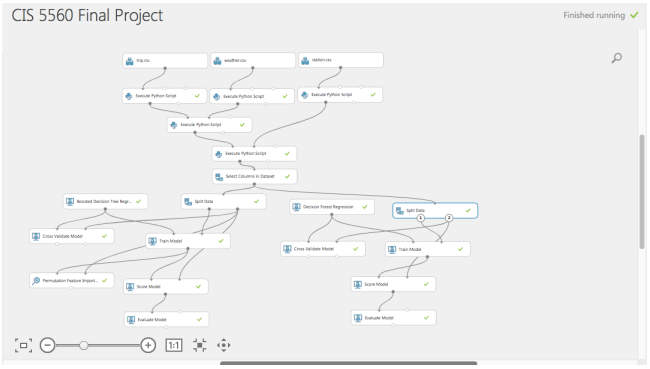


**Figure 5. Microsoft Azure Model**

The **mean_dew_point_f** is the most important feature in our table.

CIS 5560 Project › Permutation Feature Importance › Feature importance

rows 7 columns 2

view as

| Feature | Score |
|---|---|
| max_wind_Speed_mph | 3.640173 |
| max_dew_point_f | 1.760289 |
| mean_temperature_f | 1.366428 |
| wind_dir_degrees | 1.004082 |
| mean_humidity | 0.858581 |
| min_visibility_miles | 0.676148 |
| mean_dew_point_f | -0.665633 |

**Figure 6. Feature importance In azure ML**

CIS 5560 Final Project › Evaluate Model › Evaluation results

Metrics

| | |
|---|---|
| Mean Absolute Error | 26.215447 |
| Root Mean Squared Error | 32.033217 |
| Relative Absolute Error | 0.838453 |
| Relative Squared Error | 0.802699 |
| Coefficient of Determination | 0.197301 |

**Figure 7. Boosted Decision tree regression result**

- ALGORITHM USED: **BOOSTED DECISION TREE REGRESSION**
- SPLIT DATA is 70-30 and it is Randomized Split
- COD is 0.197301
- We are checking the feature important to improve the COD of the model.

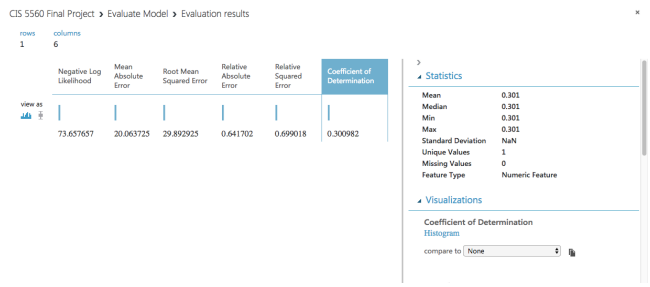


**Figure 8. Decision Forest Regression**

- ALGORITHM USED: **DECISION FOREST REGRESSION**
- COD: 0.301
- SPLIT DATA : 70 -30 is Randomized split

## 3. Summary

- We have successfully used the tools learnt in class such as databricks and microsoft Azure to built predective models and to analyse and visualize the result.
- Individually in databricks, using spark ML we come to conclusion that using Random Forest Regression algorithm gives most accurate result with cross validation of 94.81 and has the least Root mean Square error of 92.21.
- In Microsoft Azure, we have tried to improve the cofficient of Determination.Out of the two algorithms used,decision forest regression gives the best COD of 0.30 when compared to 0.197 in Boosted decision tree algorithm

## 4. GitHub URL

## References


- https://docs.microsoft.com/en-us/azure/machine-learning/studio/create-experiment
- https://forums.databricks.com/
- https://stackoverflow.com/questions/tagged/databricks
- https://docs.microsoft.com/en-us/azure/machine-learning/studio/faq