CIS5560 Term Project Tutorial

Authors: Krithy Nanaiah Atrangada; Neha Shashidhara Guli; Anupam Sahay

Instructor: [Jongwook Woo]

Date: 05/16/2018

Lab Tutorial

katrang (katrang@calstatela.edu)

sneha (sneha@calstatela.edu)

asahay (asahay@calstatela.edu)

05/16/2018

San Francisco Bay Area Bike Share Analysis

On

Data Bricks in Spark Machine Leaning

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Get data manually
- Create Spark cluster
- Train NLP system
- SQL commands to perform the analysis.

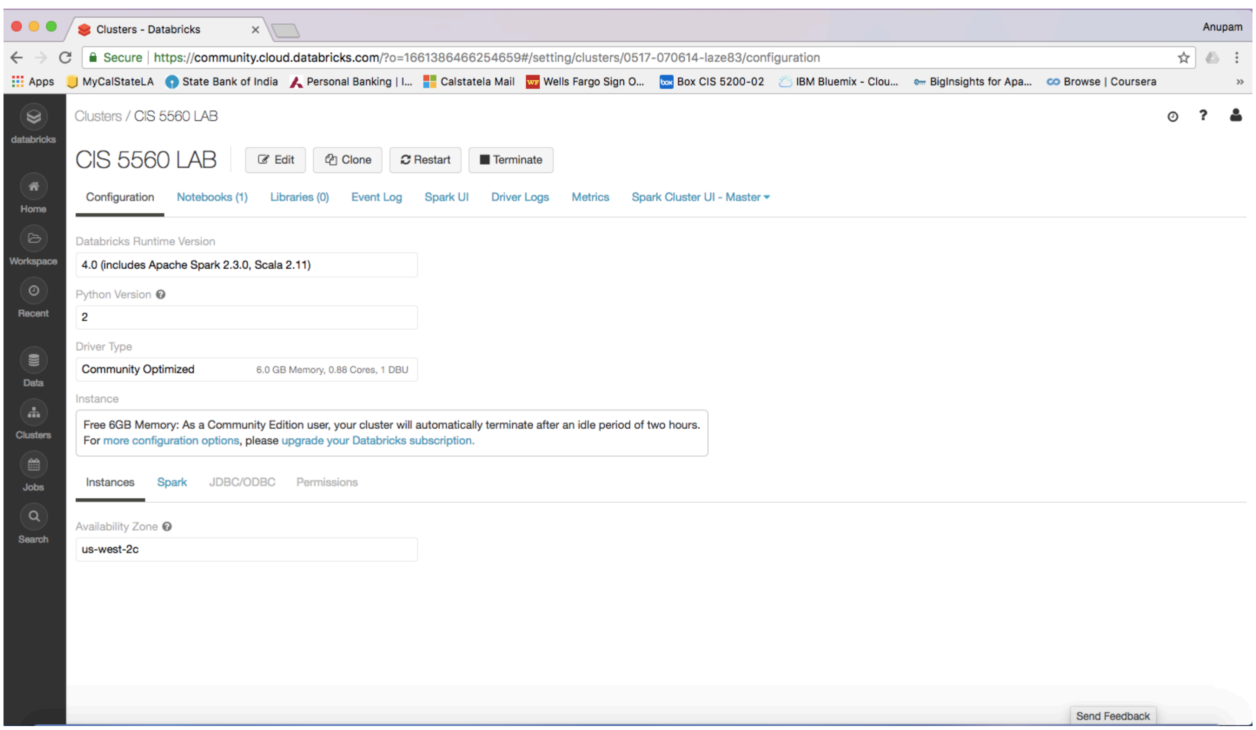
- Writing PySpark codes to develop a predictive model.
- Predicting total number of trips on a certain day using Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression.
- Visualization

Platform Spec

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

Step 1: Creating a Cluster in Data Bricks

1. This step is to create a cluster for the execution of the codes.



Properties: -

- Python Version: 2
- Driver Type: Community Optimized
- Availability Zone: us-west-2c

Step 2: Loading the Data Set in the Data Bricks

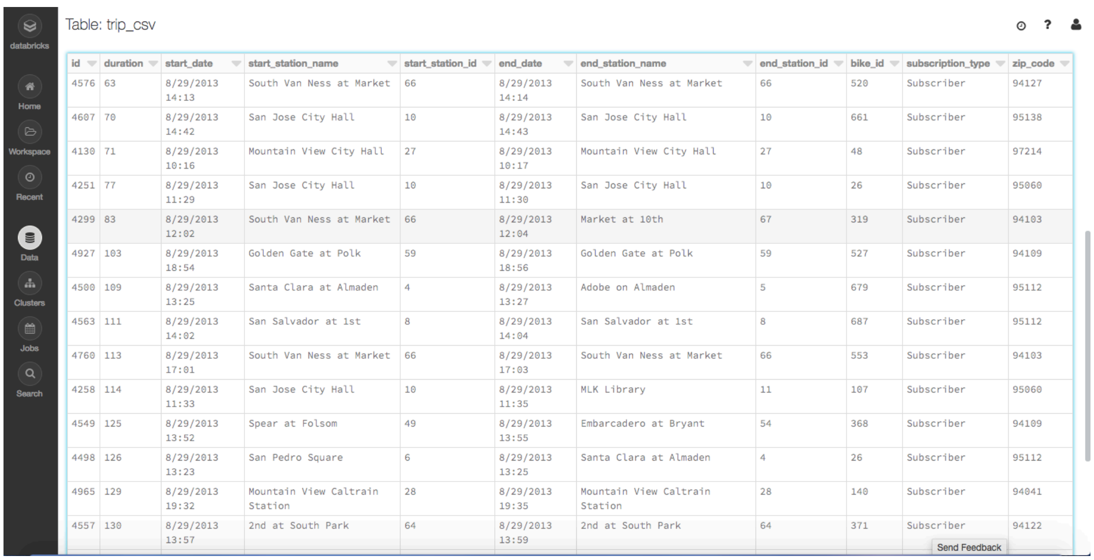


Table: trip_csv

id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name	end_station_id	bike_id	subscription_type	zip_code
4576	63	8/29/2013 14:13	South Van Ness at Market	66	8/29/2013 14:14	South Van Ness at Market	66	520	Subscriber	94127
4607	70	8/29/2013 14:42	San Jose City Hall	10	8/29/2013 14:43	San Jose City Hall	10	661	Subscriber	95138
4130	71	8/29/2013 10:16	Mountain View City Hall	27	8/29/2013 10:17	Mountain View City Hall	27	48	Subscriber	97214
4251	77	8/29/2013 11:29	San Jose City Hall	10	8/29/2013 11:30	San Jose City Hall	10	26	Subscriber	95060
4299	83	8/29/2013 12:02	South Van Ness at Market	66	8/29/2013 12:04	Market at 10th	67	319	Subscriber	94103
4927	103	8/29/2013 18:54	Golden Gate at Polk	59	8/29/2013 18:56	Golden Gate at Polk	59	527	Subscriber	94109
4500	109	8/29/2013 13:25	Santa Clara at Almaden	4	8/29/2013 13:27	Adobe on Almaden	5	679	Subscriber	95112
4563	111	8/29/2013 14:02	San Salvador at 1st	8	8/29/2013 14:04	San Salvador at 1st	8	687	Subscriber	95112
4760	113	8/29/2013 17:01	South Van Ness at Market	66	8/29/2013 17:03	South Van Ness at Market	66	553	Subscriber	94103
4258	114	8/29/2013 11:33	San Jose City Hall	10	8/29/2013 11:35	MLK Library	11	107	Subscriber	95060
4549	125	8/29/2013 13:52	Spear at Folsom	49	8/29/2013 13:55	Embarcadero at Bryant	54	368	Subscriber	94109
4498	126	8/29/2013 13:23	San Pedro Square	6	8/29/2013 13:25	Santa Clara at Almaden	4	26	Subscriber	95112
4965	129	8/29/2013 19:32	Mountain View Caltrain Station	28	8/29/2013 19:35	Mountain View Caltrain Station	28	140	Subscriber	94041
4557	130	8/29/2013 13:57	2nd at South Park	64	8/29/2013 13:59	2nd at South Park	64	371	Subscriber	94122

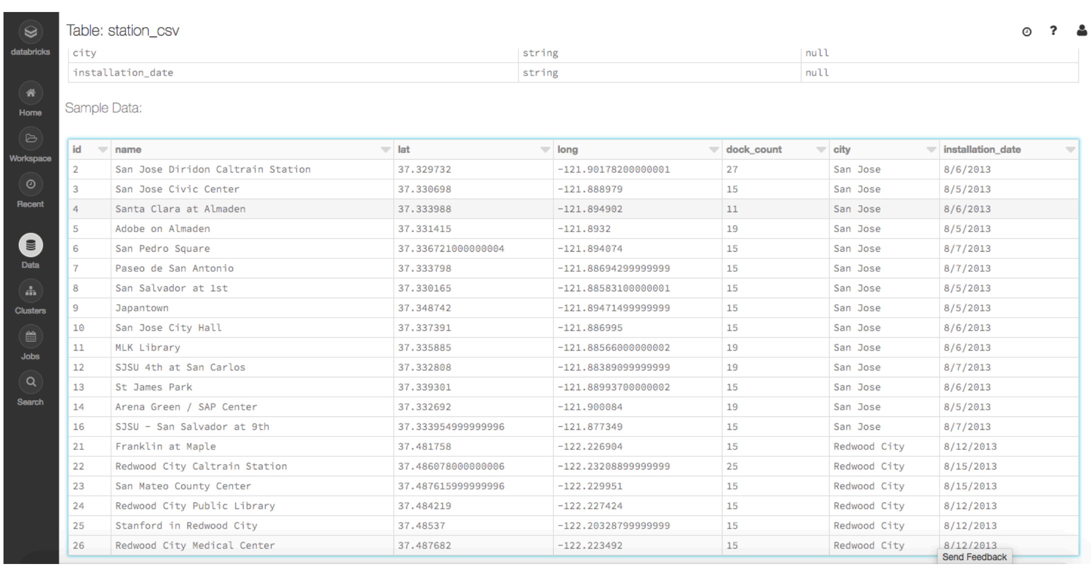
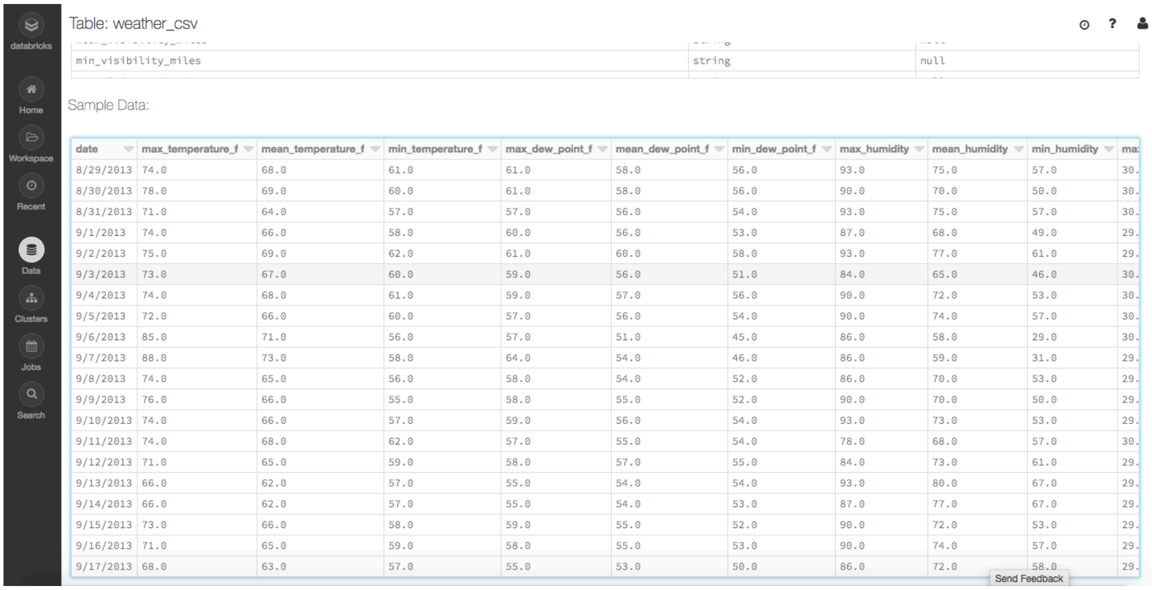


Table: station_csv

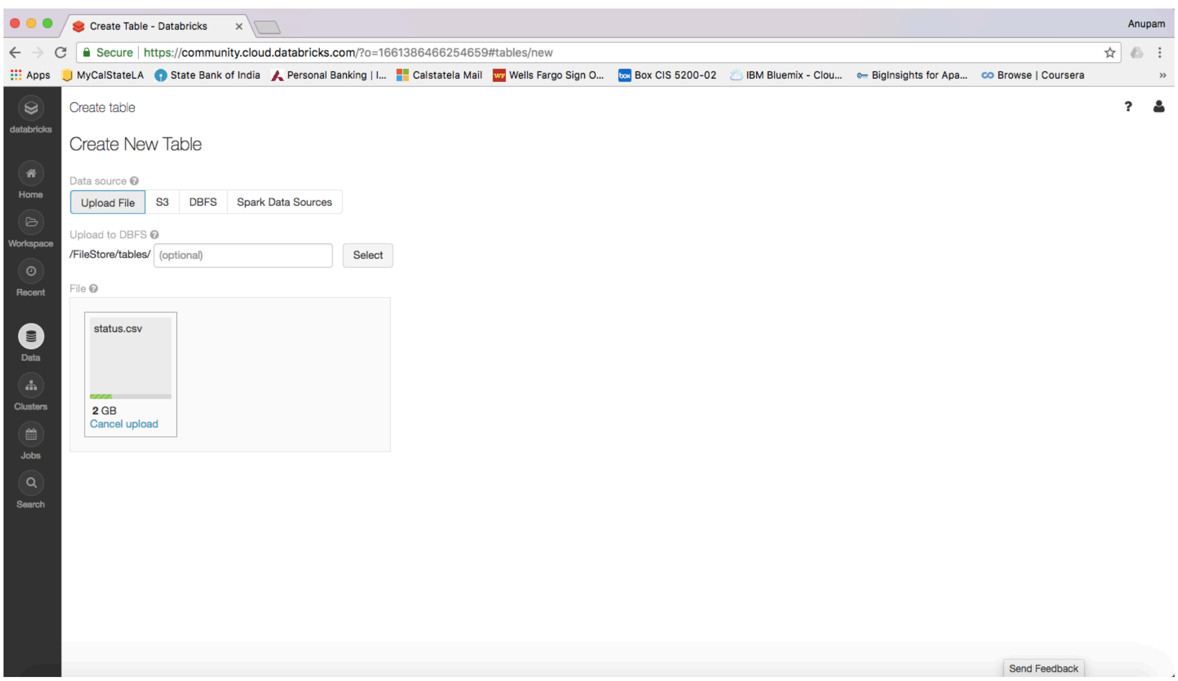
city	string	null
installation_date	string	null

Sample Data:

id	name	lat	long	dock_count	city	installation_date
2	San Jose Diridon Caltrain Station	37.329732	-121.90178200000001	27	San Jose	8/6/2013
3	San Jose Civic Center	37.330698	-121.888979	15	San Jose	8/5/2013
4	Santa Clara at Almaden	37.333988	-121.894902	11	San Jose	8/6/2013
5	Adobe on Almaden	37.331415	-121.8932	19	San Jose	8/5/2013
6	San Pedro Square	37.336721000000004	-121.894074	15	San Jose	8/7/2013
7	Paseo de San Antonio	37.333798	-121.88694299999999	15	San Jose	8/7/2013
8	San Salvador at 1st	37.330165	-121.88583100000001	15	San Jose	8/5/2013
9	Japantown	37.348742	-121.89471499999999	15	San Jose	8/5/2013
10	San Jose City Hall	37.337391	-121.886995	15	San Jose	8/6/2013
11	MLK Library	37.335885	-121.88566000000002	19	San Jose	8/6/2013
12	SJSU 4th at San Carlos	37.332808	-121.88389099999999	19	San Jose	8/7/2013
13	St James Park	37.339301	-121.88993700000002	15	San Jose	8/6/2013
14	Arena Green / SAP Center	37.332692	-121.900084	19	San Jose	8/5/2013
16	SJSU - San Salvador at 9th	37.333954999999996	-121.877349	15	San Jose	8/7/2013
21	Franklin at Maple	37.481758	-122.226904	15	Redwood City	8/12/2013
22	Redwood City Caltrain Station	37.486078000000006	-122.23208899999999	25	Redwood City	8/15/2013
23	San Mateo County Center	37.487615999999996	-122.229951	15	Redwood City	8/15/2013
24	Redwood City Public Library	37.484219	-122.227424	15	Redwood City	8/12/2013
25	Stanford in Redwood City	37.48537	-122.20328799999999	15	Redwood City	8/12/2013
26	Redwood City Medical Center	37.487682	-122.223492	15	Redwood City	8/12/2013



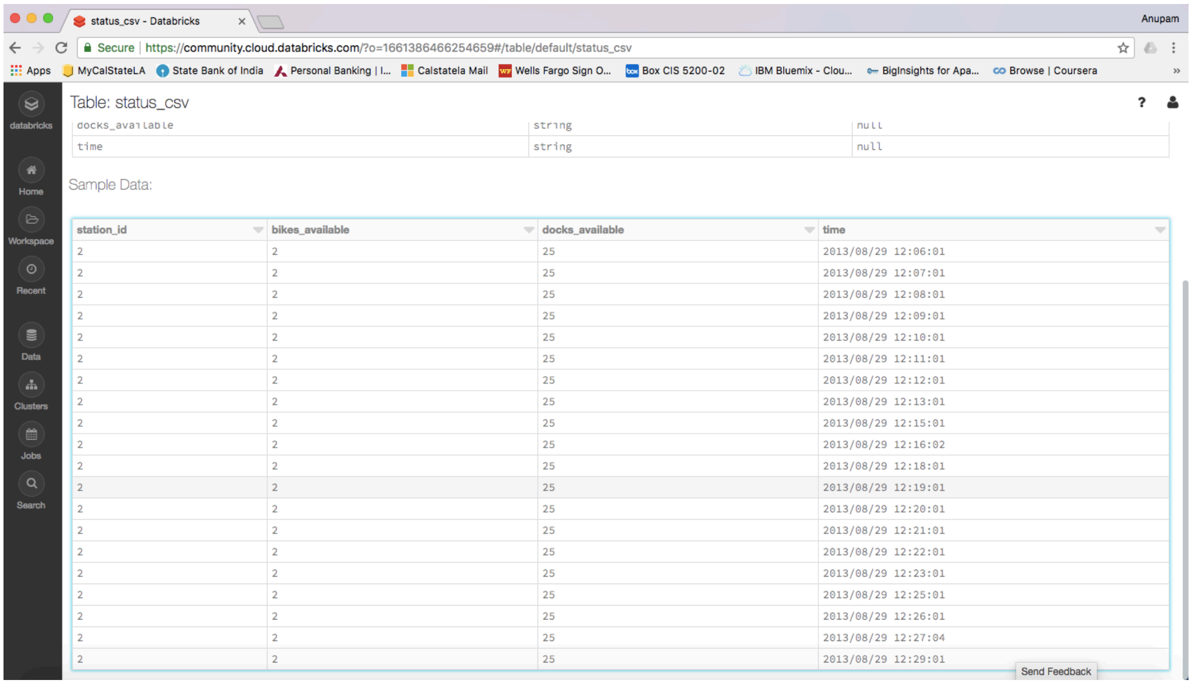
date	max_temperature_f	mean_temperature_f	min_temperature_f	max_dew_point_f	mean_dew_point_f	min_dew_point_f	max_humidity	mean_humidity	min_humidity	ma
8/29/2013	74.0	68.0	61.0	61.0	58.0	56.0	93.0	75.0	57.0	30.
8/30/2013	78.0	69.0	60.0	61.0	58.0	56.0	90.0	70.0	50.0	30.
8/31/2013	71.0	64.0	57.0	57.0	56.0	54.0	93.0	75.0	57.0	30.
9/1/2013	74.0	66.0	58.0	60.0	56.0	53.0	87.0	68.0	49.0	29.
9/2/2013	75.0	69.0	62.0	61.0	60.0	58.0	93.0	77.0	61.0	29.
9/3/2013	73.0	67.0	60.0	59.0	56.0	51.0	84.0	65.0	46.0	30.
9/4/2013	74.0	68.0	61.0	59.0	57.0	56.0	90.0	72.0	53.0	30.
9/5/2013	72.0	66.0	60.0	57.0	56.0	54.0	90.0	74.0	57.0	30.
9/6/2013	85.0	71.0	56.0	57.0	51.0	45.0	86.0	58.0	29.0	30.
9/7/2013	88.0	73.0	58.0	64.0	54.0	46.0	86.0	59.0	31.0	29.
9/8/2013	74.0	65.0	56.0	58.0	54.0	52.0	86.0	70.0	53.0	29.
9/9/2013	76.0	66.0	55.0	58.0	55.0	52.0	90.0	70.0	50.0	29.
9/10/2013	74.0	66.0	57.0	59.0	56.0	54.0	93.0	73.0	53.0	29.
9/11/2013	74.0	68.0	62.0	57.0	55.0	54.0	78.0	68.0	57.0	30.
9/12/2013	71.0	65.0	59.0	58.0	57.0	55.0	84.0	73.0	61.0	29.
9/13/2013	66.0	62.0	57.0	55.0	54.0	54.0	93.0	80.0	67.0	29.
9/14/2013	66.0	62.0	57.0	55.0	54.0	53.0	87.0	77.0	67.0	29.
9/15/2013	73.0	66.0	58.0	59.0	55.0	52.0	90.0	72.0	53.0	29.
9/16/2013	71.0	65.0	59.0	58.0	55.0	53.0	90.0	74.0	57.0	29.
9/17/2013	68.0	63.0	57.0	55.0	53.0	50.0	86.0	72.0	58.0	29.



STEPS:

- First, we click on Data on left hand side.
- Then we upload the file from the local disk on to the data bricks
- Once the file is uploaded click on create table with UI.

- Select the cluster which we created. Then click on Preview Table
- In Specify Table attributes check first row is header. Then click on create table.



Step 3: Train Natural Language Processing

This step explains the codes which are used in the Data Bricks for execution.

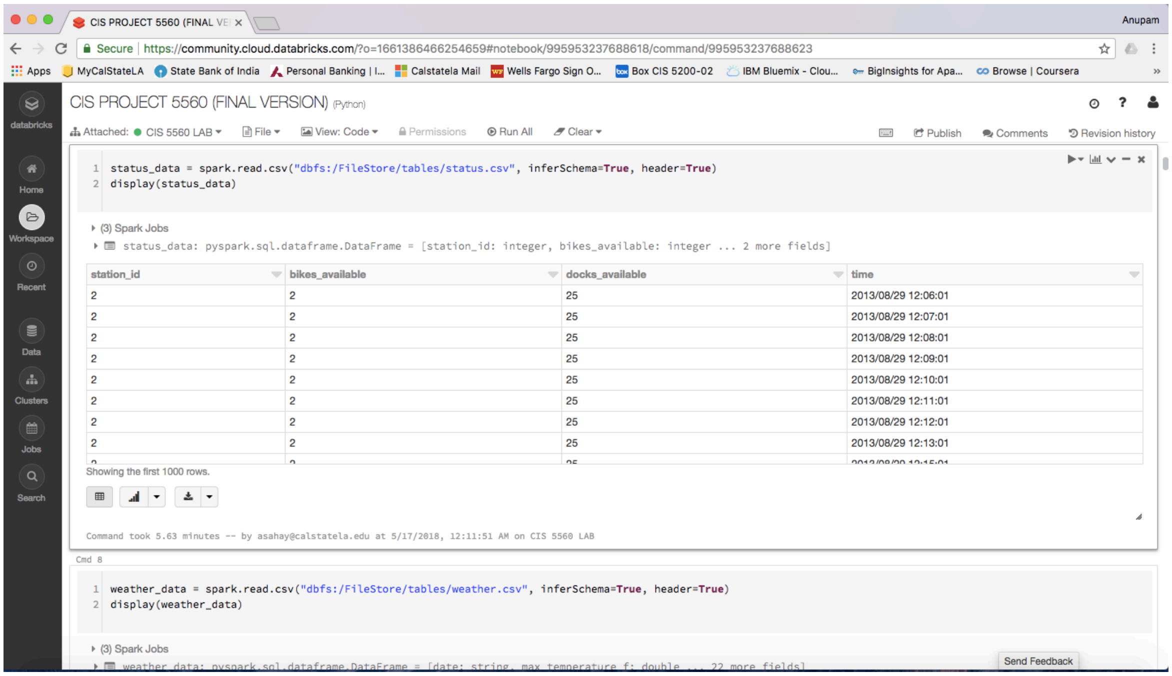
%fs ls /FileStore/tables/status.csv

%fs ls /FileStore/tables/weather.csv

%fs ls /FileStore/tables/trip.csv

%fs ls /FileStore/tables/staion.csv

- In this step we load the data in the data bricks file system.
- We then display the file. In the data bricks to check if all the files are loaded from the source.



- When all the data is loaded on the data bricks, we then check if any of the tables contains any null values.
- We check for the null values for all the dataset.

from pyspark.sql.functions import isnan, count, when

status_data.select([count(when(isnan(c), c)).alias(c) for c in status_data.columns]).show()

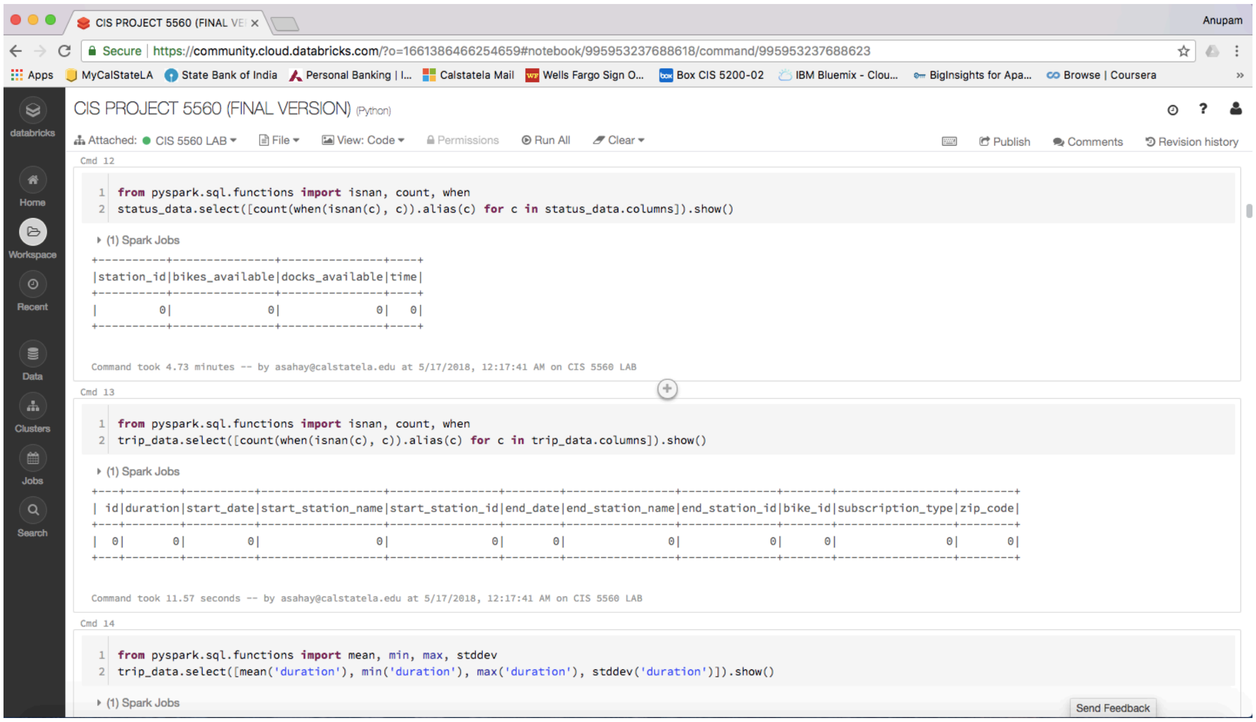
from pyspark.sql.functions import isnan, count, when

trip_data.select([count(when(isnan(c), c)).alias(c) for c in trip_data.columns]).show()

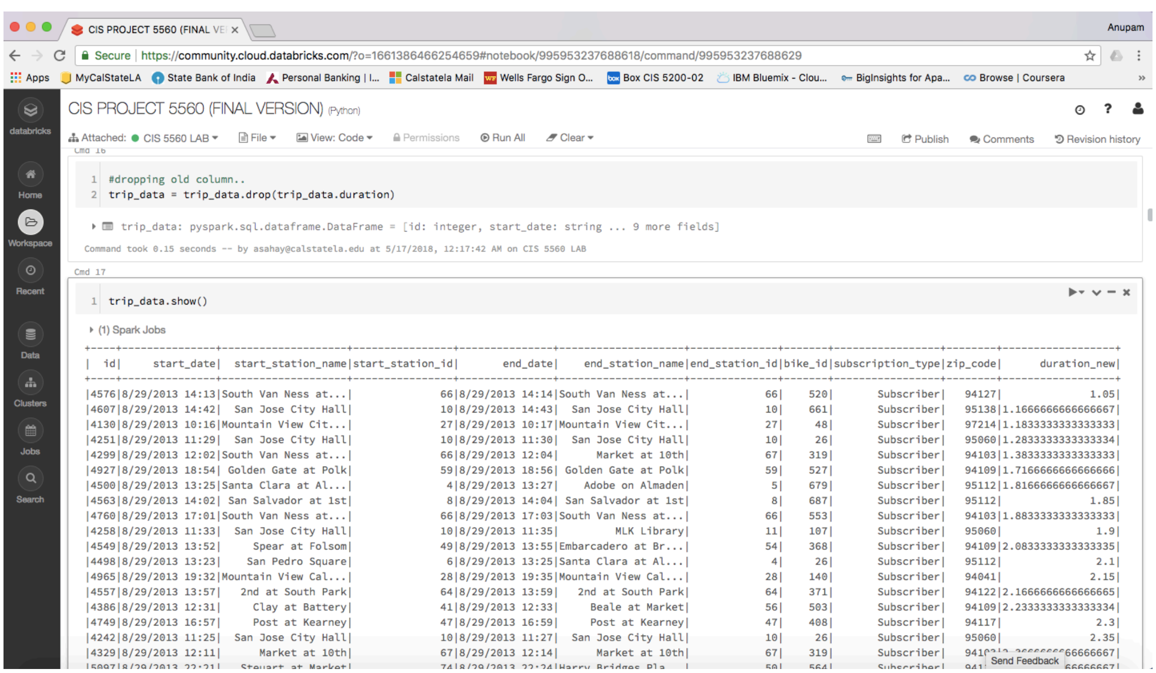
- We then check for the max,min,mean,median for the trip_data.

from pyspark.sql.functions import mean, min, max, stddev

trip_data.select([mean('duration'), min('duration'), max('duration'), stddev('duration')]).show()

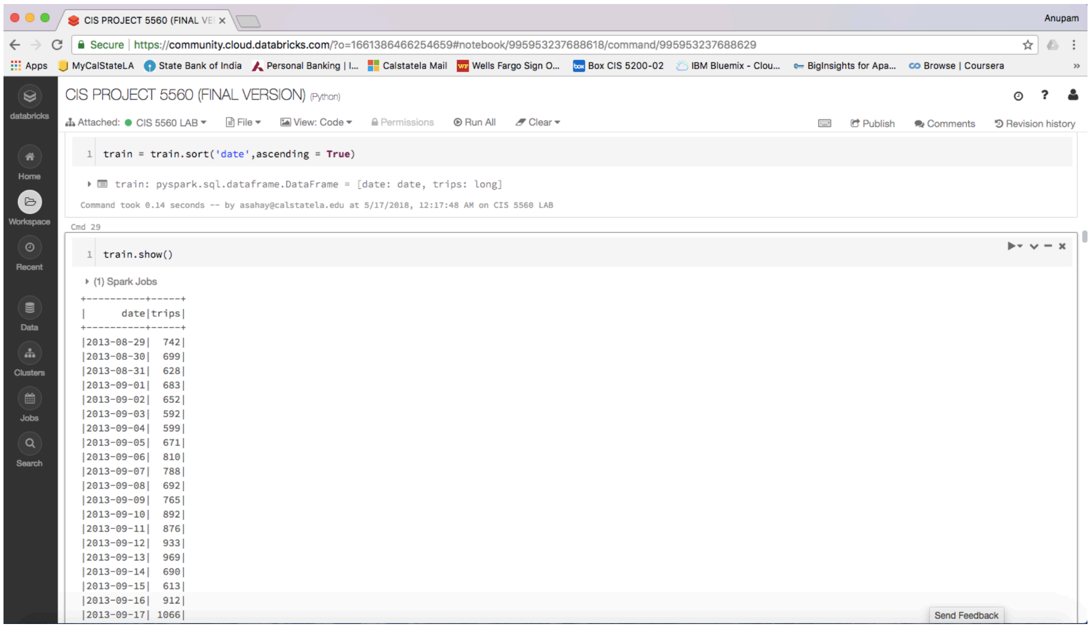


- We then create a new duration in which we convert all the seconds to minutes and we drop the old columns.
- We then again check for the max, min,mean,median for the trip_data.



- Then we convert it into data time so that it can be manipulated easily.

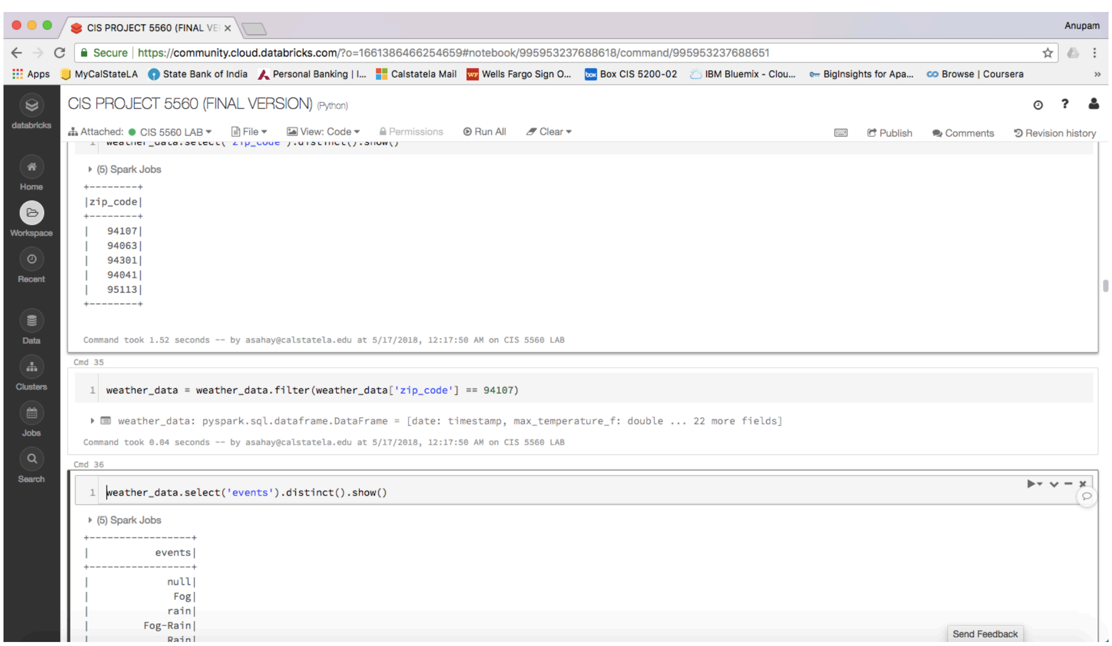
- We then look for the distinct trip and we sort them in ascending order



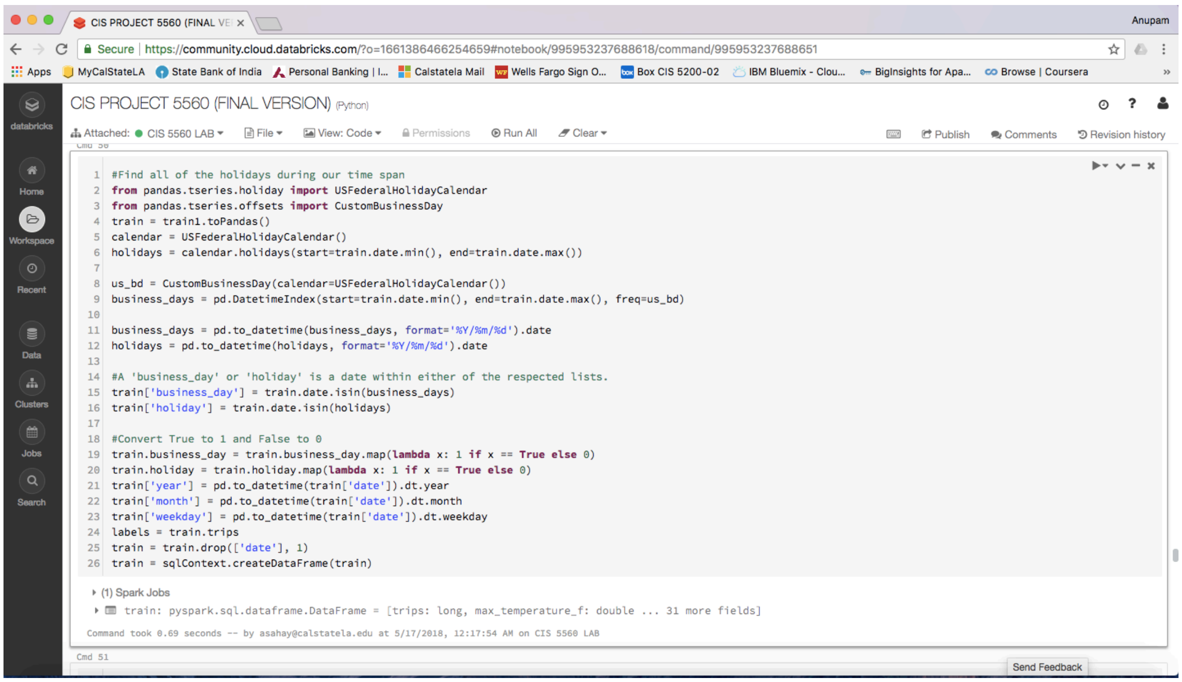
- We do similar transformation for weather data and we look out for distinct zip codes and weather events which will affect our prediction.

weather_data.select('zip_code').distinct().show()

weather_data.select('events').distinct().show()

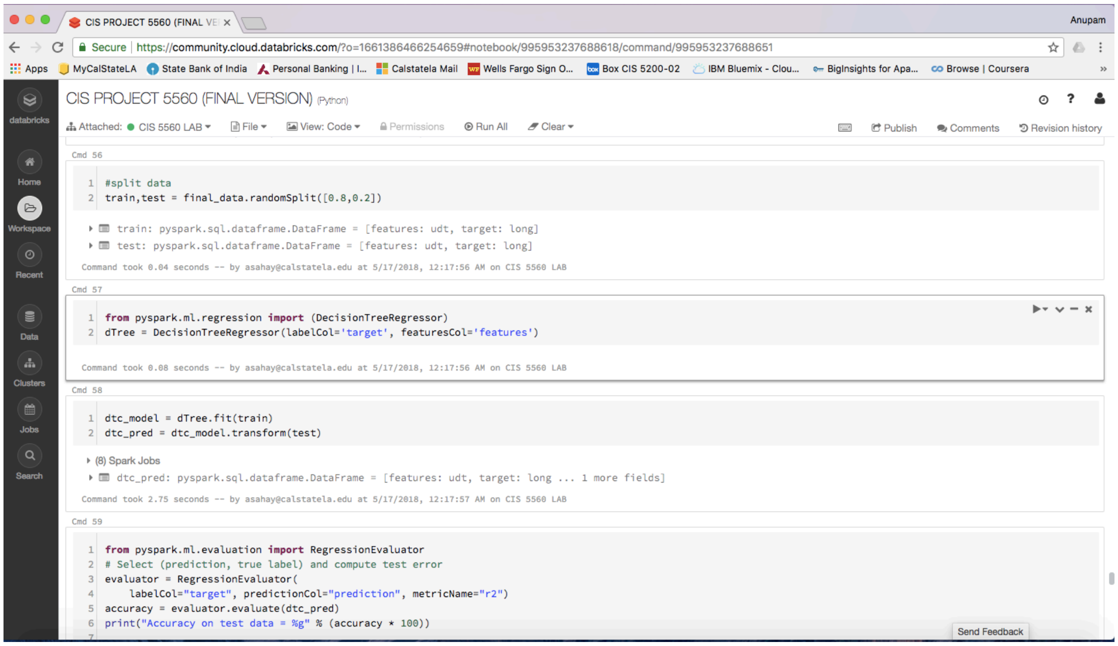


- We import US Federal Holiday Calendar in our data and we treat the holidays/weekens as 0 and the working days/business days as 1.
- We make True=1 and False=0. This is required to find out the total number of holidays during the time span.
- We again convert all these parameters in the format of “date”, “month”, and “year”.



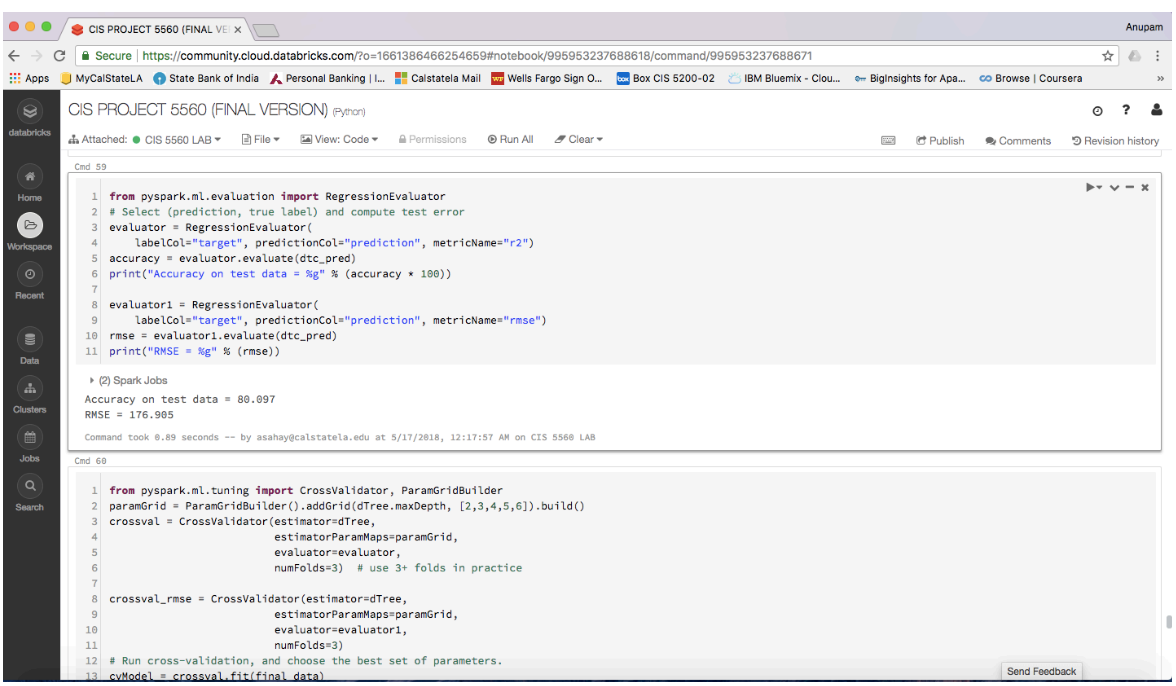
- We split the data as train and test in the ratio of 0.8, 0.2 respectively.

train,test = final_data.randomSplit([0.8,0.2])

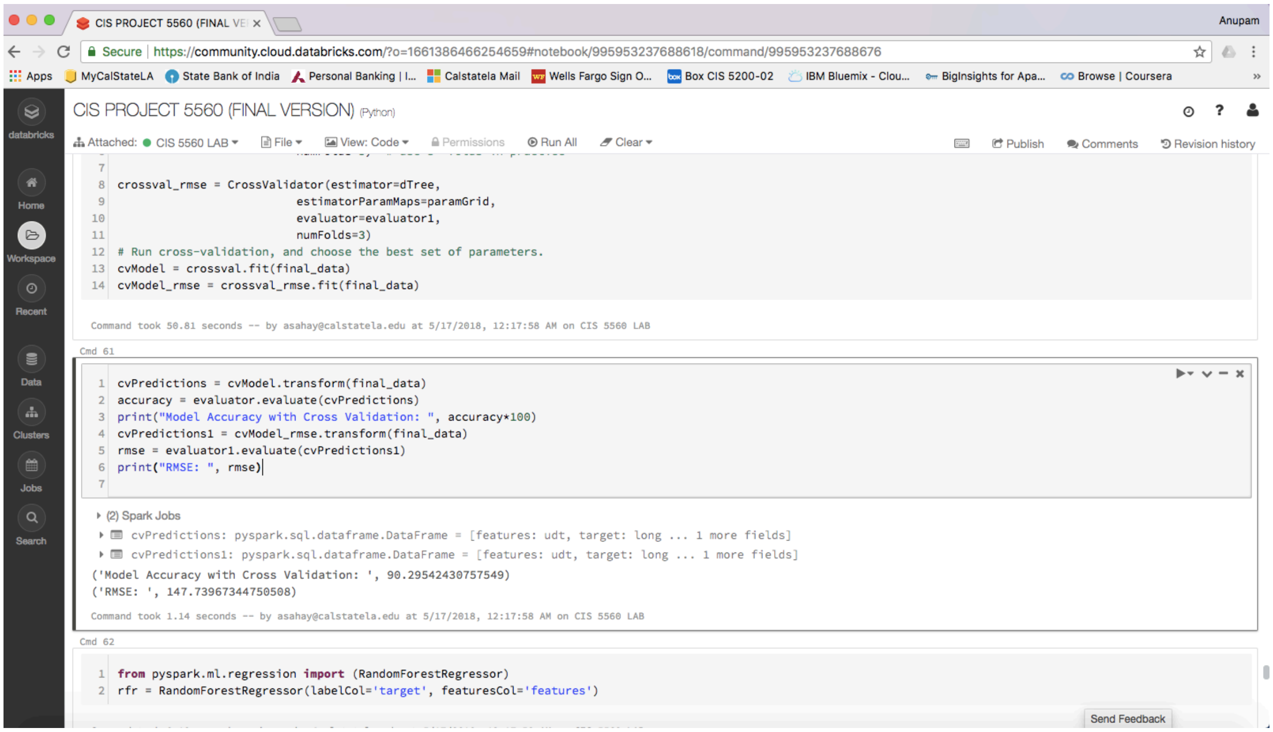


Step 4: Model 1

- After the preparation of the data we use our first algorithm to build our model.
- We use Decision Tree Regression Algorithm.
- We find out the accuracy on the test data and the root mean square error.



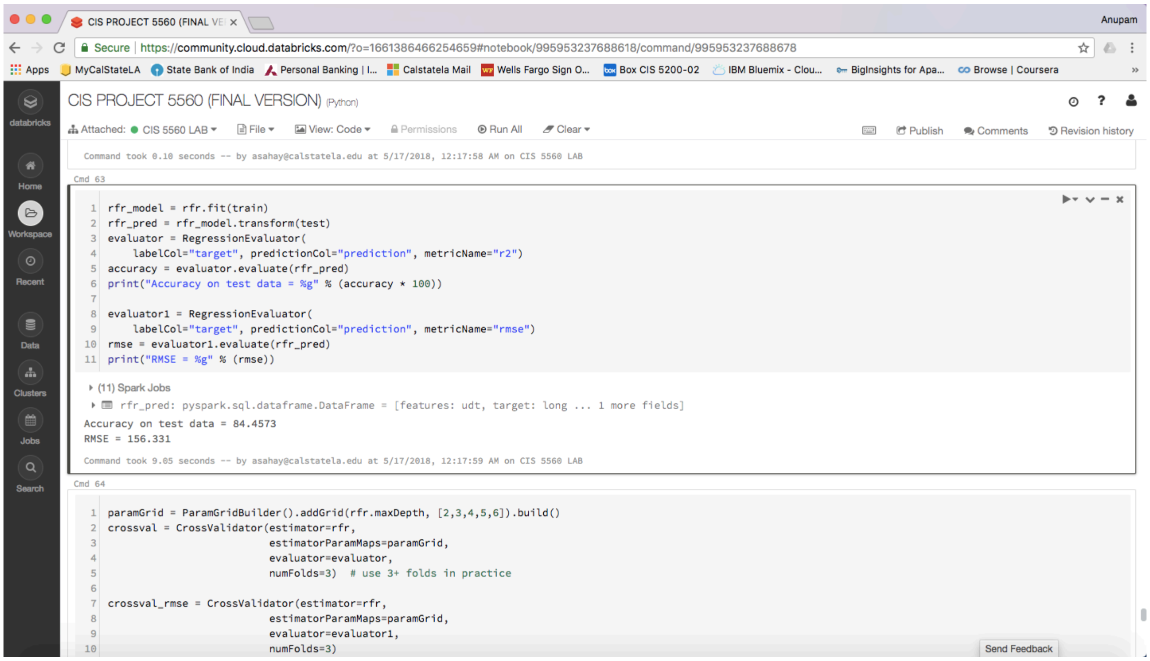
- In this we can see that the accuracy on the test data is 80.097 and the RMSE is 176.905.
- After using the cross- validation where K =3 the accuracy is improved and the RMSE is reduced.



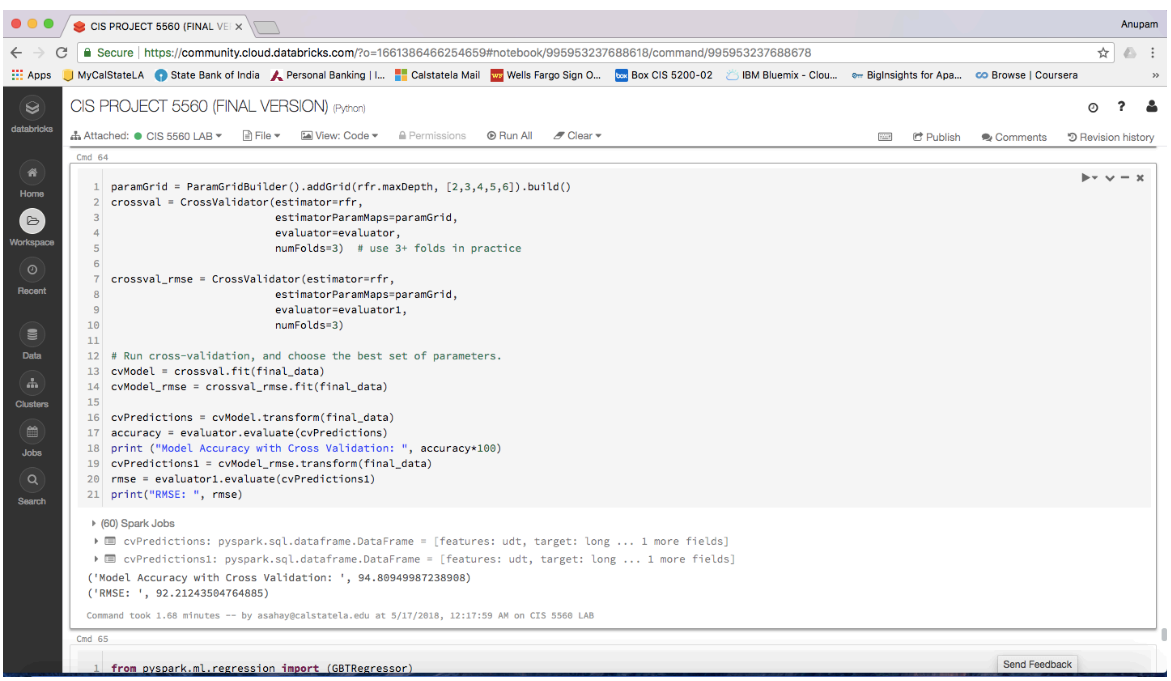

```
('Model Accuracy with Cross Validation: ', 90.29542430757549)
('RMSE: ', 147.73967344750508)
```

Step 5: Model 2

- We use Random Forest Regression Algorithm.
- We find out the accuracy on the test data and the root mean square error.
- The accuracy on the test data is 84.45 and the RMSE is 156.331
- After using the cross- validation where K =3 the accuracy is improved and the RMSE is reduced significantly.



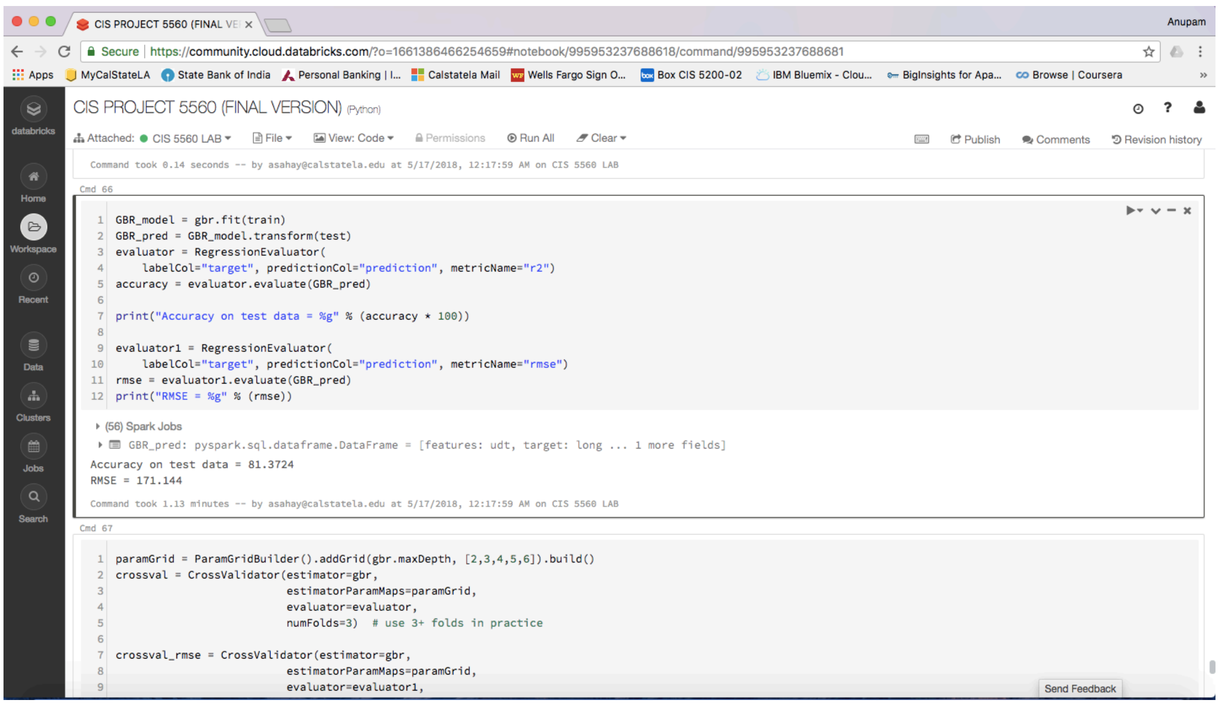
After the cross-validation process: -




```
('Model Accuracy with Cross Validation: ', 94.80949987238908)
('RMSE: ', 92.21243504764885)
```

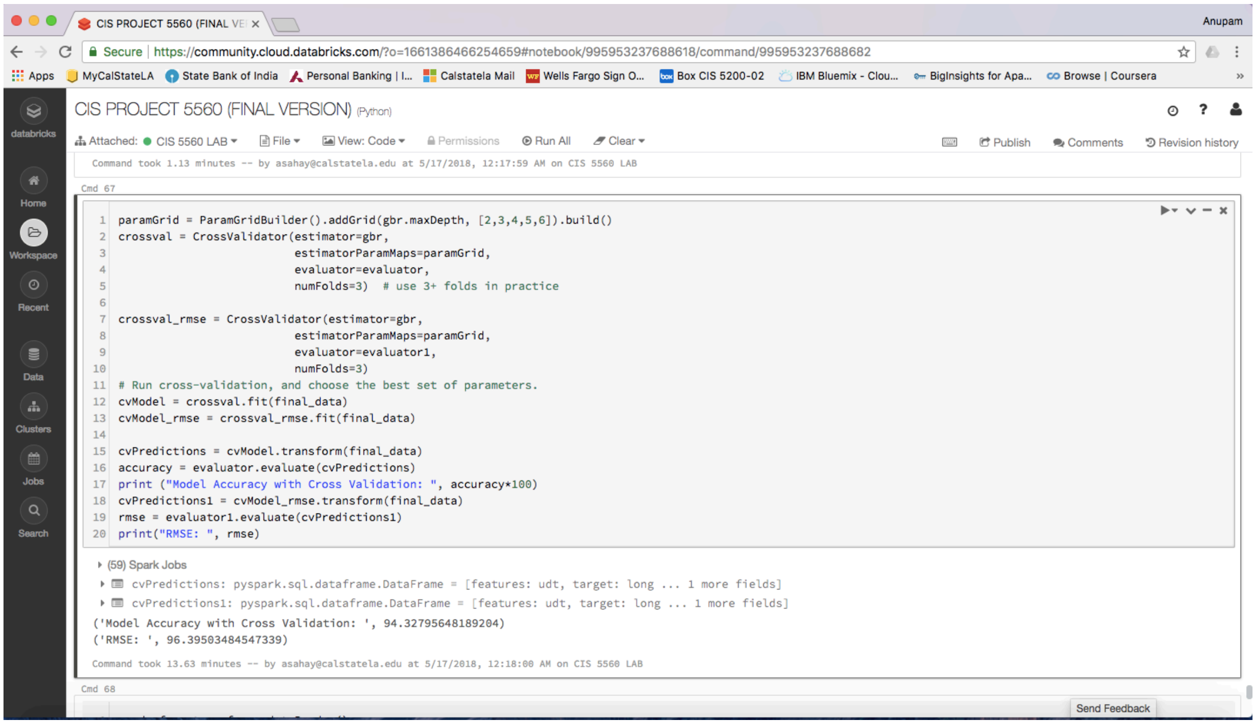
Step 6: Model 3

- We use Gradient Boosting Regression Algorithm.
- We find out the accuracy on the test data and the root mean square error.
- The accuracy on the test data is 81.3724 and the RMSE is 171.144
- After using the cross- validation where K =3 the accuracy is improved and the RMSE is reduced significantly.



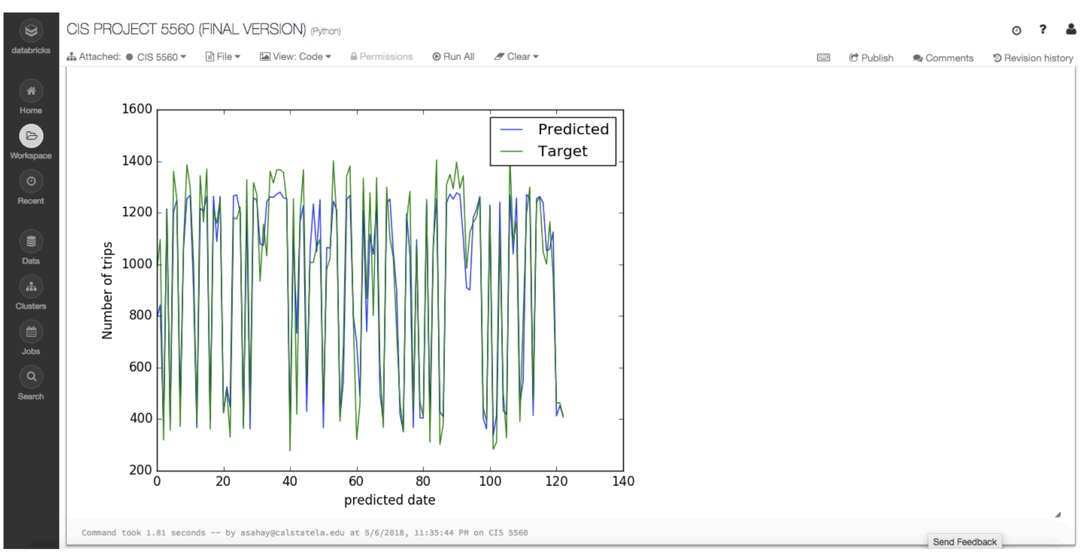
After implementation of the cross-validation process: -

```
('Model Accuracy with Cross Validation: ', 94.32795648189204)
('RMSE: ', 96.39503484547339)
```



- From above three models we can find out that the Random forest regression is the best model for the prediction.
- We make the visualization of the model by using the graphs.

Step 7: Visualization (Graph)



- In this visualization we are showing the target number of trips which takes place on a particular day, and number of trips predicted by using Random Forest Regression Algorithm.
- These models a give a prediction of number of trips. We do think that we have made a good model.

References: -

1. https://forums.databricks.com/
2. https://stackoverflow.com/questions/tagged/databricks