Project: Capstone Project 2:-

Project name: - Citi Bike– Repricing case study and Twitter Sentiment analysis for real time reputation management

Student Name: - Jitendra Agarwal

Course: - Springboard cohort Jan2 2018

Summary: - Citi Bike is the largest bike share program in us, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens and Jersey City. It was designed for quick trips with convenience in mind, and it's a fun and affordable way to get around town. Everyone knows that bike sharing is the answer to many environmental and urban transportation issues, yet it's not mainstream in US. I am being asked by the senior executive team at Citi bike to use data science techniques to recommend 3 key action item to increase the company's' business.

Problem Statement: -

- How is the performance of unit economics per trip, per bike, and per station over the months?
- is there is any statistical way to find the reason for a decline in trips observed few times last quarter?
- How many bikes should be placed at a station so that customer always finds a bike when he needs?

Project Goals: -

- Exploratory analysis to find meaningful patterns in data
- Predict the right number of bike to stationed at a station
- User Sentiment analysis from Twitter and identify most common customer issues and how to address them.

Data Source: -

Data is provided by Citi bike for academic project purpose here: -
https://www.citibikenyc.com/system-data

This data is provided according to the policy
https://www.citibikenyc.com/data-sharing-policy

Raw data downloaded from: -
https://s3.amazonaws.com/tripdata/index.html

Download following 3 zip file and unzip them

201710-citibike-tripdata.csv.zip
201711-citibike-tripdata.csv.zip
201712-citibike-tripdata.csv.zip

1. Pre Processing and Data Wrangling

Code:- ./capstone-project2-Jitendra_part_1

- Input for python notebook is 3 csv files with raw data: -

```
201710-citibike-tripdata.csv
201711-citibike-tripdata.csv
201712-citibike-tripdata.csv
```

1. Read Data from csv • Divide data in 2 part of monthly subscriber and Day pass holder
2. Convert birth year to Int and calculate user age
3. Filter data for user age > 100 as an outlier
4. Rename columns to remove spaces from their name
5. Add a unique trip id column for each row as a unique key.
6. Extract Date and hour from timestamp column
7. Determine if a trip is Free or paid, and calculate paid units and paid amount for each trip
8. Create a new DF with unique list of station list
9. Drop column which are not needed to reduce DF size
10. Get Weather data from csv and store in Data Frame.
11. Calculate Monthly unit metrics such as
    a. Total distance o Average distance per trip
    b. Average time per trip
    c. Average distance per bike
    d. Average time per bike
    e. Average trip per bike
    f. Average Revenue per bike
    g. Average revenue per trip
12. Calculate monthly revenue by different streams

- Output of python notebook is saved in following csv files: -

1.	../data/df_summ_DV.csv: -	Monthly summary of trips and unit economics
2.	../data/df_temp_DV.csv: -	Temperature for 3 months of Quarter 4 2017
3.	../data/df_trip_bymonth_DV.csv: -	Summary by month for Data Visualization
4.	../data/df_daily_trips_ML.csv:-	Daily trip summary for machine learning
5.	../data/df_cust_DV.csv:-	All trips by pass holder after cleanup
6.	../data/df_sub_DV.csv:-	All trips by Monthly Subscribers after cleanup

2. Data Visualization

Code:- ./capstone-project2-Jitendra_part_2

- Input for this python notebook is 5 csv files with cleaned up data: -
    - ../data/df_summ_DV.csv: - Monthly summary of trips and unit economics
    - ../data/df_temp_DV.csv: - Temperature for 3 months of Quarter 4 2017
    - ../data/df_trip_bymonth_DV.csv: - Summary by month for Data Visualization
    - ../data/df_daily_trips_ML.csv:- Daily trip summary for machine learning
    - ../data/df_cust_DV.csv:- All trips by pass holder after cleanup
    - ../data/df_sub_DV.csv:- All trips by Monthly Subscribers after cleanup

1. Plot Key unit metrics per trip and per bike for each month.
2. Plot No of trips happened each day of Q4 and plot it with the average temp for each day
3. Plot trip distribution by hour of the day in entire Q4 for subscriber and customer
4. Plot Trip Duration of each trip and show the threshold of free trip limit of 45 minutes for Subscribers
5. Trip Origination Distribution by Station id: where are all trip starting from
6. Paid Trip Distribution by Day/Time of a week: When are the paid trips happening
7. Paid Trip Distribution by Day/Time of a week: When are the paid trips happening
8. Plot Bike Trip Distribution by Gender and Age
9. Plot Trip counts by Bike Ids
10. Plot Bike Trip Distribution by Gender, Age and paid v/s Free Trips
11. Track and Plot Bike Movement In Peak Hours
12. Create a Heat map to show which station is busy against the hour of the day.

3. Machine Learning Models

Code:- ./capstone-project2-Jitendra_part_3

- Input for this python notebook is 1 csv files with cleaned up data created in part 1: -

../data/df_daily_trips_ML.csv

Following steps were completed to create an optimal machine learning model.

1. Heat map of the covariance between features
2. ols model to predict no of trips from a station for a given day
3. Plot Regression between No of trips and other features
4. Create LinearRegression model
5. cross validation and Recursive feature elimination
6. Plot residuals
7. Create a Leverage Plot
8. Remove outliers
9. Recreate the test and train split after removing outliers. New data frame with features as 'density_score', 'start_id', 'start_day', 'AVGT'
10. Performing a evaluation of multiple models to see which has best outcome.
11. Tuning and Evaluation of Random forest repressors and Extra tree repressors
12. Create class of function to try these models with various parameter options.
13. Show the summary of score of all models / parameter combination in sorted order.
14. Create and Fit final tuned model for predicting no of trips from a station based on feature

Final outcome-

Based on Machine Learning exercise we can predict the no of trips from a station with 87% accuracy.

The no of trips can be predicted by providing the values of Station id, Average temperature, station density score and Day of the week.

Most important feature for prediction is Station id with 50% importance.

Random Forest Repressors was used as best Machine learning model

After tuning, train data prediction accuracy was 97% and test data prediction accuracy was 87%.

Citi Bike trip data volume is huge and it has high variance in distribution as not all places have an equal footprint of bike riders.

Naturalized mean Square error value was only 0.06770262

4. twitter user sentiment analysis using NLP

Code:- ./capstone-project2-Jitendra_part_4

Following steps were completed to create an optimal machine learning model for sentiment analysis

1. Access twitter using Tweepy API
2. Get twitter data for has tags #CitiBike','#citibike', '#bikenyc and save in csv file.
3. Create new columns in data frame for tweets and calculate values.
4. Plot No of Tweets by Day and Tweet Lengths along time:
5. Plot sentiment score along with length of tweet / review
6. Plot the tweet count based on the sentiment outcome
7. Removing stop words and punctuation for proper bigram creation
8. Cleanup tweet using The BeautifulSoup Package.
9. Plot word cloud
10. Plot most common words in tweet
11. Initialize the CountVectorizer object, which is scikit-learn's bag of words tool.
12. CountVectorizer converts a collection of text documents to a matrix of token counts.
13. Fit vectorizer to create train and test set by bag of words
14. Fit using TfidfTransformer to create train and test set by bag of words
15. Fit the Random forest classifer forest to the training set, using the bag of words as features and the sentiment labels as labels
16. Create Pipeline to try and evaluate different model and parameter combinations
17. Show the summary of score of all models / parameter combination in sorted order.

Final outcome: -

45% tweets are with positive sentiment, 42% with neutral and 13% tweets with negative sentiment 78% yelp reviews are having negative sentiment

The combined sentiment is 39% positive and is towards negative due to high negative Yelp reviews.

NLP machine learning model used Logistic regression classifier and accuracy was 81%

Reference: -

- Trip data source:- https://www.citibikenyc.com/system-data campaign (CitiBike)
- Weather data:- http://w2.weather.gov/climate/
- yelp data:- Yelp public reviews open source web scraping.
- Tweet Data:- twitter API
- Inspirations 1:- https://github.com/toddwschneider/nyc-citibike-data\
- Inspirations 2:- http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/
- Inspirations 3:- http://toddwschneider.com/posts/taxi-vs-citi-bike-nyc/
- Inspirations 4:- https://joomik.github.io/Housing/