BELLEVUE UNIVERSITY

Project Report

DSC630 - Predictive Analytics

Gangadhar Dhulipala
Anirban Pal
Raghu Raman Nanduri
August 10, 2019

Contents

1 EXECUTIVE SUMMARY

Demand management is very important for any transportation-based company. For the bike ride sharing use case that we considered, planning and preparing for the demands of the biking enthusiastic customers, is paramount to be a successful business. With that business objective we want to use machine learning approach for understanding the usage pattern and predicting number of bikes required for each station, clustering the stations based on the trips and predicting passengers membership type. The goal of the project was to improve the profitability with enhanced rider experience by accommodating customers demands in a timely manner. We analyzed The effect of factors such as weather, time of day, day of week, membership type on bike traffic, stuffing and profitability.

From the analysis, it has been seen that 'Monthly pass' and 'Walk up' are the two most popular and 'Flex pass' is the least popular membership type based on number of trips, 'Walk up' having the most ride duration.

Quite interestingly, the most popular station with maximum outgoing trips, 'Ocean Front Walk Navy', is inactive today.

During October and December of 2018, there has been significant surge in rental volume. The first quarter of either year(2018, 2019) has been consistently low in traffic. Winter could be the reason for the volume.

Looking at the trip distribution during the week, Sunday seems to be the most popular day for biking.

2 TECHNICAL REPORT

Data Report

We combined and used three different data sets for the purpose of this project. These were

- Bike trip information (130k records)
- Station information (140+ records)
- Weather information (National Oceanic and Atmospheric Administration - NOAA)

Trip Data:

trip_id: Locally unique integer that identifies the trip

duration: Length of trip in minutes

start_time: The date/time when the trip began, presented in ISO 8601 format in local time

end_time: The date/time when the trip ended, presented in ISO 8601 format in local time

start_station: The station ID where the trip originated (for station name and more information on each station see the Station Table)

start_lat: The latitude of the station where the trip originated

start_lon: The longitude of the station where the trip originated

end_station: The station ID where the trip terminated (for station name and more information on each station see the Station Table)

end_lat: The latitude of the station where the trip terminated

end_lon: The longitude of the station where the trip terminated

bike_id: Locally unique integer that identifies the bike

plan_duration: The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up)

trip_route_category: "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips

passholder_type: The name of the passholder's plan

bike_type: The kind of bike used on the trip, including standard pedal-powered bikes, electric assist bikes, or smart bikes.zduration: Length of trip in minutes

start_time: The date/time when the trip began, presented in ISO 8601 format in local time

end_time: The date/time when the trip ended, presented in ISO 8601 format in local time

start_station: The station ID where the trip originated (for station name and more information on each station see the Station Table)

start_lat: The latitude of the station where the trip originated

start_lon: The longitude of the station where the trip originated

end_station: The station ID where the trip terminated (for station name and more information on each station see the Station Table)

end_lat: The latitude of the station where the trip terminated

end_lon: The longitude of the station where the trip terminated

bik_id: Locally unique integer that identifies the bike

plan_duration: The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up)

trip_route_category: "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips

passholder_type: The name of the passholder's plan

bike_type: The kind of bike used on the trip, including standard pedal-powered bikes, electric assist bikes, or smart bikes.

Station Data:

Station ID: Unique integer that identifies the station (this is the same ID used in the Trips and Station Status data)

Station Name: The public name of the station. "Virtual Station" is used by staff to check in or check out a bike remotely for a special event or in a situation in which a bike could not otherwise be checked in or out to a station. Go live date: The date that the station was first available

Region: The municipality or area where a station is located, includes DTLA (Downtown LA), Pasadena, Port of LA, Venice

Status: "Active" for stations available or "Inactive" for stations that are not available as of the latest update

Weather Data: We only considered the temperature component of the weather report since we could not get complete information about the other components from the weather website. All the temperatures are in °F.

DATE: Date of the weather reading.

TMIN: Minimum temperature of the day

TAVG: Average temperature of the day

TMAX: Maximum temperature of the day

Exploratory Data Analysis:

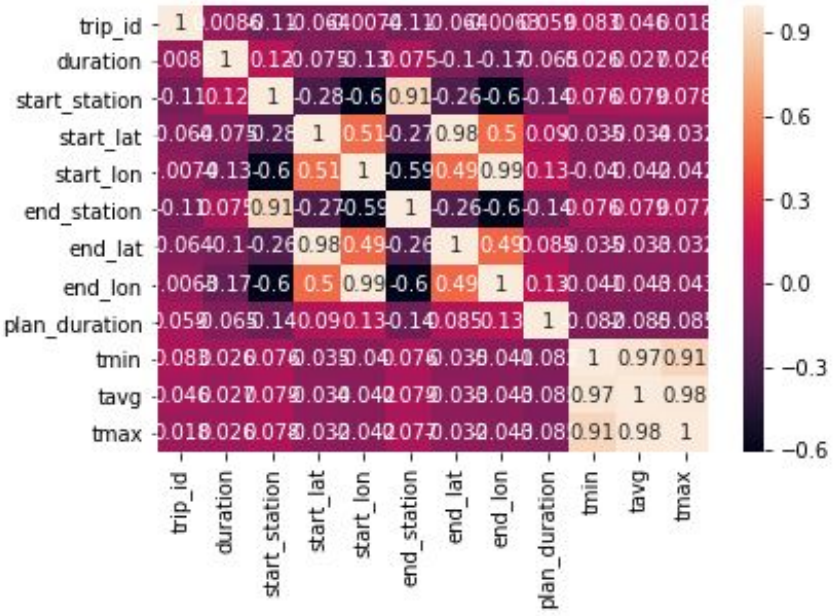

Figure 2.1: Correlation Matrix

Some of the attributes show strong correlation e.g. ‘end_lat‘ vs ‘start_lat‘, ‘end_lon‘ vs ‘start_lon‘ etc. But those are mostly because of the trips those started and ended at the same station. We were looking for correlation between trip attributes and weather attributes but no such strong correlation exist. Therefore the data needs further investigation.

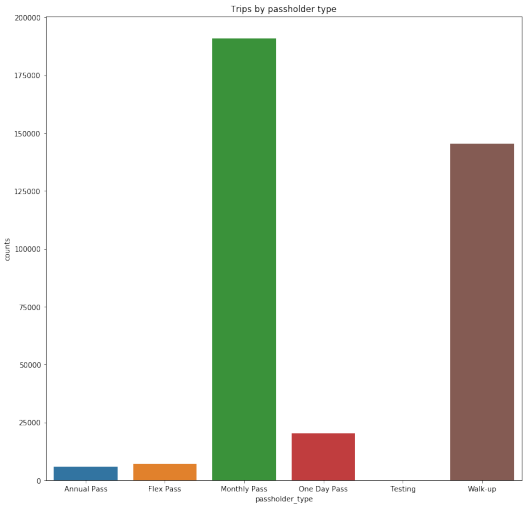

Figure 2.2: Trips by passholder type

Monthly pass and Walk-up are usually the most common trip types. One Day Pass, Flex Pass and Annual Pass are not used that much when compared to the Monthly Pass.

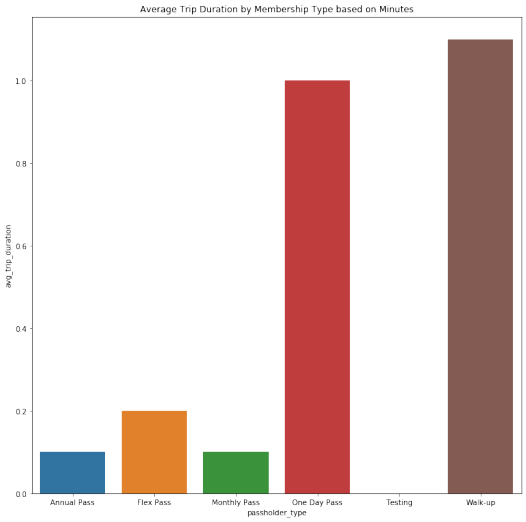

Figure 2.3: Average Trip Duration by Membership Type based on Minutes

Trip duration of one day pass and walk ups are much greater than monthly.

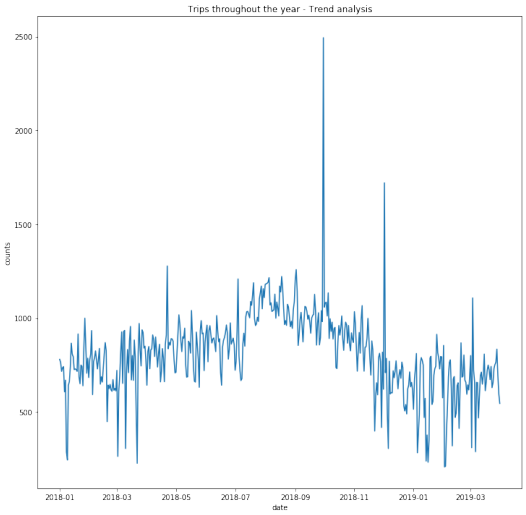

Figure 2.4: Trip trend

October and December seems to be the busiest month.

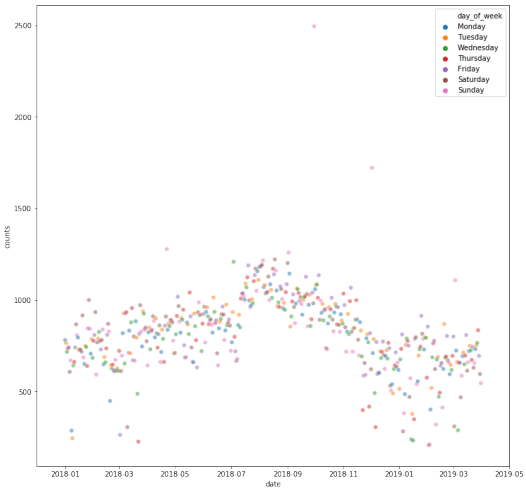

Figure 2.5: Trip trend

Sunday seems to be the busiest day of the week.

Predictive Modelling:

- **Problem 1:**

  Problem Statement:

The first business objective we want to achieve is better demand management of the number of bikes required to meet the customer needs in a timely manner. Proper planning and preparing to withstand the demands of the bikes is paramount to be a successful bike ride sharing business. To achieve that business objective we want to use machine learning approach for understanding the usage pattern and predicting number of bikes required for each station.

Feature Engineering Selection:

As part of feature engineering, we have created some derived variables that are more useful in our regression analysis. We create the following variables: 1) Dummy variables for categorical variables âĂŞ Station_id status 2) Derived variables 'outgoingbikes' 'incomingbikes' based on trips that start from the station and end at the station respectively. 3) Derived Holiday flag based on dates of national holidays and marking Sundays and SaturdayâĂŹs as holidays as well to study the impact of the holidays on the bike sharing trips No feature selection was done as part of this regression analysis, as we know the correlated variables, we have manually discarded those variables for example removed âĂŞ station name (as we already have created dummy variables for station Id) date field (after converting it into numerical month year). For date, we instead used weekday by deriving it from date and created a holiday flag to study impact of holiday on the bike share rides.

Building Models:

The dependent variables that are we are trying to predict here to solve this business objective are numerical and continuous in nature. But the dependent variables are not linear in nature with respective to the predictors, so we have considered to go with non-linear regression analysis. After data preparation activities including data cleansing, merging and feature engineering we have trained data with both Decision Tree Regressor and Random Forest Regressor.

Model Evaluation:

Since we are dealing with regression problem here, we have used Mean Absolute Value, Mean Square Value to verify the error in predictions based on both training and test dataset. We have also used R âĂŞ Square values to see how much of a variance in dependent variables is explained by the models. Also plotted residuals vs predicted values to see whether there is any pattern in the residuals. The lack of a clear-cut pattern confirmed that

Decision Tree Regressor Metrics:

Mean Absolute Error is : 4.278630579680873

Mean Squared Error : 81.57856998586144

R-square value : 0.14232187678190478

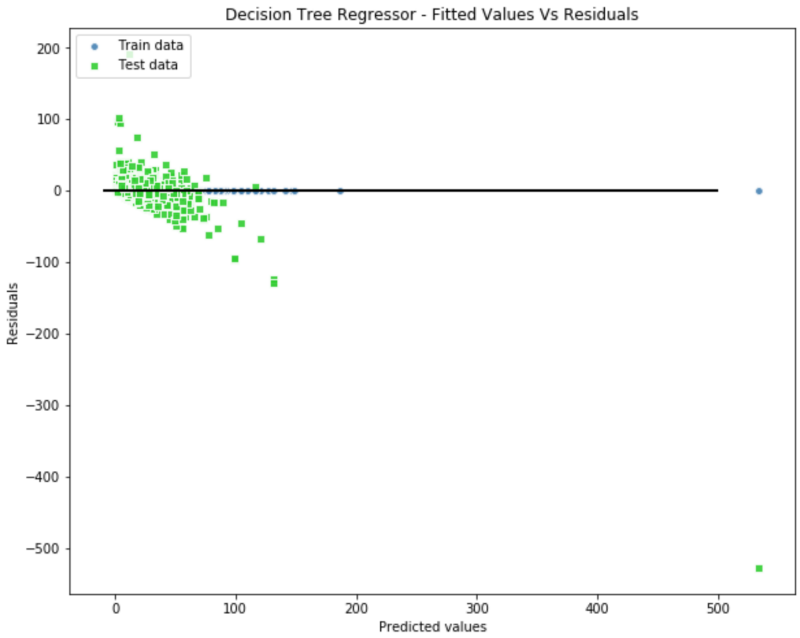

Figure 2.6:

Random Forest Regressor Metrics:

Mean Absolute Error is : 0.3322798597501033

Mean Squared Error : 0.34051502104901665

R-square value : 0.6594849789509833

- **Problem 2:**

Problem Statement

As the second problem statement, we wanted to cluster the stations based on ‘totaltrips‘ and ‘tavg‘ (average temperature, using unsupervised learning.

Approach

For clustering, we decided to exclude the status column. Since status is specific to the time of acquiring the data, dropping record based on status or using status as a feature might lead to data loss or inaccuracy. We also created three groupings of high (trips > 100), medium (trips between 50-100) and low traffic (trip < 50), based on the number of total trips for the stations. We used *k-means* clustering (3 folds) for this model. It was observed that increasing the number of folds more than 5 is making the cluster radius

very specific and creating a potential problem of oversampling. *k-means* can be applied as

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c^j \right\|^2 \tag{2.1}$$

where, J = Objective function k = Number of clusters n = Number of cases x = Case i c = Centroid for cluster j Clustering has surfaced that day of the week and temperature has significance in trip count.

Model

K-Means clustering from scikit learn has been used to create the unsupervised grouping. Since there are three groupings of stations, k=3 has been used. The features have been scaled to improve performance.

Evaluation

The k-means run has generated accuracy of 93%.

- **Problem 3:**

  Problem Statement:

  As the third problem statement, one of the question we are asking about the data set is that- given some attributes related to a trip, whether we will be able to predict whether a certain passenger belong to a monthly pass holder type or just a regular walk-up type customer. Able to predict the pass holder type given his few trips information adds huge value to the company as it can use this information to understand about the future revenue of the company that will be generated from the monthly pass holders.

  Approach

  The ‘trip_merged‘ data set comprises of our comprehensive data set. However, the catch was that, there are a total number of 336236 rows and 25 attributes in the data set. Since the predictor variable is the pass holder type, we first looked at the different type of pass holders present in the data set. To begin with there are 6 different types of pass holders and out of which the ‘monthly pass‘ and the ‘walk-up‘ are the two major ones that outnumber the other ones. Hence, data set is selected in such a way that all the instances correspond to these two type of pass holders.

  In addition, majority of the attributes like, ‘trip_id‘, ‘start_lat‘, ‘start_lon‘ etc. were removed from the analysis as these does not have any role in the overall prediction. By doing such data prepossessing, the number of relevant attributes were reduced to 19 including all the dummies that were generated for the categorical variables.

Model

The model used for this study is a Random Forest classifier from the scikit learn package. The data set was further split into test and training data set before it was fed to the algorithm. The classifier object was further applied to the test data set.

Evaluation

The model was evaluated by generating a confusion matrix as well as calculating the area under the curve.

Accuracy

The Random Forest Classification model was a very good to the data and it gave an accuracy of about 99%.

3 ACKOWLEDGEMENT

[1] Guo Y, Zhou J, Wu Y, Li Z (2017) Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China. PLoS ONE 12(9): e0185100 https://doi.org/10.1371/journal.pone.0185100

[2] Rani M., Vyas O.P. (2017) Smart Bike Sharing System to Make the City Even Smarter. In: Bhatia S., Mishra K., Tiwari S., Singh V. (eds) Advances in Computer and Computational Sciences. Advances in Intelligent Systems and Computing, vol 553. Springer, Singapore.

4 REFERENCES

[1] Metro Ride Share data, retrieved from https://bikeshare.metro.net/about/data/

[2] Weather data, retrieved from https://www.noaa.gov/