From Data to Insights: The Blueprint for SF Bay Area Bicycle Trip Business

Vinothkumar Ganeshan
M1 Master's in Machine Learning and Data Mining
University Jean Monnet
vinothkumar.ganeshan@etu.univ-st-etienne.fr

Abstract

The project aims at mining the different attributes of the bicycle usage by the people of San Francisco bay area to propose the new business insights based on the data. It gives a lot of interesting information about the trip data of users to new business models and any preventive measures to be taken to solve future problems. The effective usage of PCA, clustering techniques and different data mining algorithms like Apriori and Eclat algorithms are used effectively to mine the great insight about the living of people to propose a new business model can be constructed with the help of R and Rattle.

1 Introduction

Bike trips are a repository of bike usage data in different cities of bay area fig 1.1 by categories of people during the period between August 2013 to August 2015. It consists of major five cities like Mountain View, Palo Alto, Redwood City, San Francisco and San Jose.

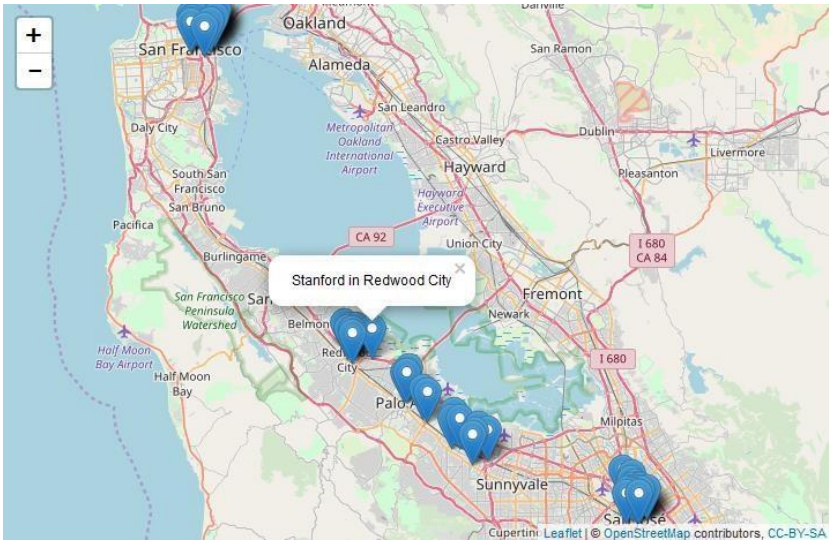

Fig 1.1 Bike stations in Bay Area

As simple as a daily task, the person may be a commuter, student, or tourist, rides the cycle as part of his/her journey. The data is in a very structured format for efficient knowledge discovery process and mining more information is more interesting too as a treasurer for a business model which has given great insights. Let's dive deep into the information to get more about it.

2 Problem Understanding:

The problem with the obtained dataset is to understand the people pattern behind the usage of bicycles over the period of time. We are more focused to understand and find the business insights of this data which would help to improve the service and to come up with new business ideas to promote and increase the usage among the users.

Business insights:

- Best possible way to improve the trip count.
- Distributing the traffic by setting up new stations at peak stations.
- Efficient bike management and maintenance
- Managing the maintenance without disturbing the bike trips
- Increasing the subscriber count by improving the service.
- Investment Planning
- Bike Maintenance
- People preference spot to improve new places.
- Association rules over the Subscriber Vs Customer.

2.1 Data structure:

For our study, we are using the kaggle bike trip SQLite[1] database which has the following properties. The dataset has a lot of features composed of different purposes like trip, station, weather and status. The features of the dataset has its own value with respect to its position.

Firstly, *Station* dataset has the information about its station id, an available number of docks, location, city and installation date of the stations.

The status dataset consists of bikes and docks available with respect to the station id with the limit of the one-minute interval of all the stations.

The trip dataset is the golden treasure of this dataset to understand the people's behaviour. It consists of people journey information like from, to station with respect to date and time.

Finally, *Weather* dataset gives the information about various parameters of weather with respect to the date and time like temperature, humidity, wind speed etc.…

Being a simple dataset, it comes up with a lot of interesting facts and became the hidden treasury for new or existing business model to have unforeseen findings. It helps to understand the usage of bikes with respect to multiple factors to gain information.

[1] https://www.kaggle.com/benhamner/sf-bay-area-bike-share/data

We cared more about the trip information in the dataset which has the valuable information hidden which involves people actions, thus the analysis on trip, station, and weather datasets are given high preference to understand the more business insights.

3 Data Understanding:

In the process of data understanding[2], the goal is to gain general insights about the data that will potentially be helpful for the further steps in the data mining process. By having the neutral point of view, the raw dataset to be modelled helps to have a very good knowledge about it before the deployment. As, we are more concerned about the trip dataset, where we consider the duration of trips has *outliers* where it has 80% of the trips were under 10 minutes. Thus, removing the outliers helps to get the more relevant reliable data model to be built. The occurrence of same trip model helps decide people has the high preference over the system. The classification of the dataset based on the Weekend and weekdays also has the high potential to see insight about people preference based on the trips. By the end of this phase, we came to have a conclusion by having a very good understanding of the data based on its format, model and data preparation model for the business insights of problem understanding in the node.

4 Preprocessing:

The dataset has been preprocessed to clean, aggregate, summarize and visualize the data as to be prepared for knowledge mining as a data frame. To implement the data mining algorithms, the data has been converted to a set of transactions by considering each trip as a transaction. The data was preprocessed based on the need for knowledge discovery to figure out by converting it into weekday and weekends.

The preprocess the dataset based on the need of the day to '%m/%d/%Y %H:%M' to respective day and date for effective knowledge mining of the dataset. The data has been split into train and test data to understand the prediction models to fit and to gain information about the accuracy of the model. To have a very good model of understanding, outliers such as trips greater than 6 hours were truncated to understand the true behaviour of the model. The trip data based on data model were separated as weekend and weekdays as well.

5 Apriori Deployment:

In this project, we more focused towards the trip data to understand how it was distributed over the period of time. I wanted to use an influential algorithm to mine the frequent itemset where the Apriori algorithm uses a "Bottom-up" approach, where the frequent subsets are extended one item at a time. The ability of Apriori Algorithm to process a huge dataset, I was looking forward to understanding the trips and how much it was related based on the Subscription type. As an Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

The data were preprocessed to remove any outliers such as trips having a duration greater than 6 hours which corresponds to less than 0.5% of the total dataset. It is devised to operate on a database

[2] https://link.springer.com/chapter/10.1007/978-1-84882-260-3_4

containing a lot of transactions, by considering every trip as a transaction in the database. Data Mining, also known as Knowledge Discovery in Databases(KDD), to find anomalies, correlations, patterns, and trends to predict outcomes. The support and confidence interval of the Apriori algorithms are used to find the frequent itemsets of transactions as a trip. Analyzing the frequent itemsets of gives interesting facts about the transactions. The tuning parameters of the Apriori algorithm are effectively used to mine the frequent itemsets of the transactions in the trip id.

Subscriber →. 2
Customer → 1

The ability to apply the confidence and minimum support for every transaction (i.e every trip) over the database, the *lift* which is the likelihood of bike and station id to occur whenever there is a trip makes sense for the business model to understand the estimation of trips before the occurrence of the trips.

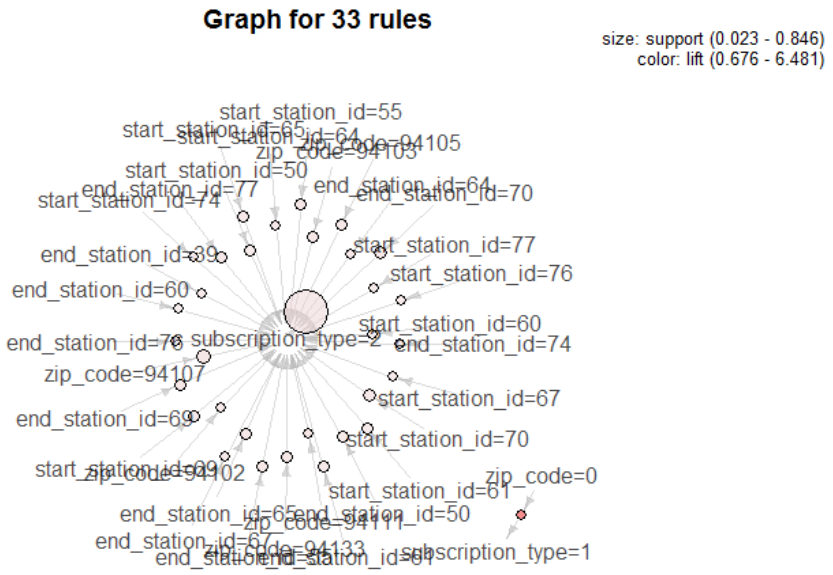

Fig 5.1 Apriori Mining on the dataset

The transactional dataset reveals the information of how each trip works. It gives the interesting fact that the Subscriber trips has highest confidence interval which reveals the maximum information gain about the trips. The rules for formed from the support 0.22, confidence 0.5 and the lift 0.676 which reveals the information of subscriber has associated with the stations and the city.

From the obtained association rules, it is clear to see that the Subscription_type =2 (Subscriber) has the higher number of trips than the Customer (Subscription_type = 1).

It also interests to find the correlation in the trip attributes such as city Zip_code, Start_station_id, end_station_id.

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5   0.022      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 14739
```

Business insights derived from this method :

- The count of Customer is very low which needs to take care of attractive offers to make them a subscriber.
- Investment planning strategy would be focused on high traffic areas, where the station id 69, 70, 74 has high traffic which needs to take care to distribute the traffic.
- The station id 20, has negligible or under usage count which can be changed to another place.
- The Zip_code 94133, 94102 which belongs to San Francisco reveals that the San Francisco city has the high number of users than other cities.
- People preferred stations are observed where it could be utilised to improve the service of that station to perform better.

Thus this algorithm gives the maximum information as it could be used in the business process to understand it better.

5.1 Do bike stations are stable for people:

The major difference among stations to act as source and destination trips in fig 5.1 reveal the information about the usage of trips stations. The huge difference states the inadequate number of bicycle available over the period of time, where ***Station 70*** has the major difference. Some of the stations between are used in a very negligible size where removing or creating new stations among the most used station would help to distribute the traffic among the stations.

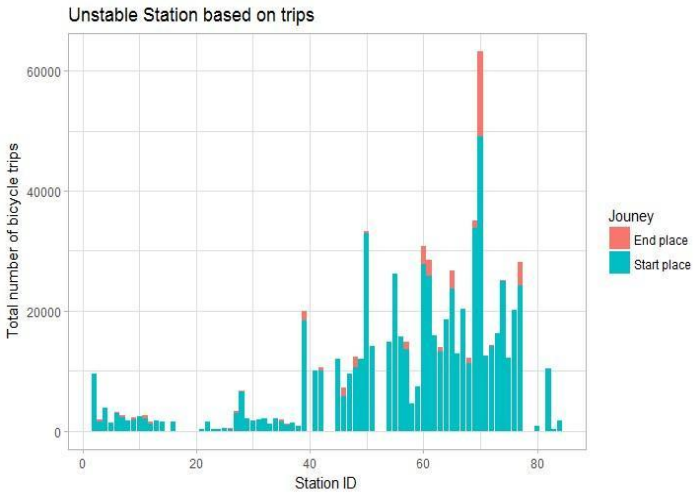

Fig 5.1 Unstable station analysis

The priority preference to be given to that unstable station to have number of cycles and distribute the traffic among the nearby stations. This mined information is more useful for *investment planning* for the next level of processing. The station ids which have a very negligible amount of trips over the period are not effectively used so that the number of bikes and the docks can be reduced to put into the place where it was really needed so much, which makes the life easier for people. This information helps to understand that the preferred station for many people is Station 70 to end their journey but less number of people start from over there. It makes the **imbalance** in the count of trips, such that there would be more bikes deposited in the station 70 which makes that there won't be enough bikes to start from other stations. Thus utilising this information helps to understand that in some stations there is not enough number of bikes to start their journey. As a business insight to distribute the bikes in station 70 to nearby stations to improve the count of trips for the people.

5.2 Clustering of Trips:

Clustering as multiple objective functions which is used to group objects in the same order that which it belongs. The trips were clustered fig 5.2 to understand the usage of trips over the period. The efficient clustering of trips based on the day and the count of trips helps to understand that next day trip which it belongs to. The clustering helps to forecast the needs of the day and the expected count of trips helps to understand the usage of bikes over the period of time. **K means** algorithm to classify efficiently to which cluster the count of the trip belongs to over the period of time, helps to understand the property of that trips and to be prepared for their need.

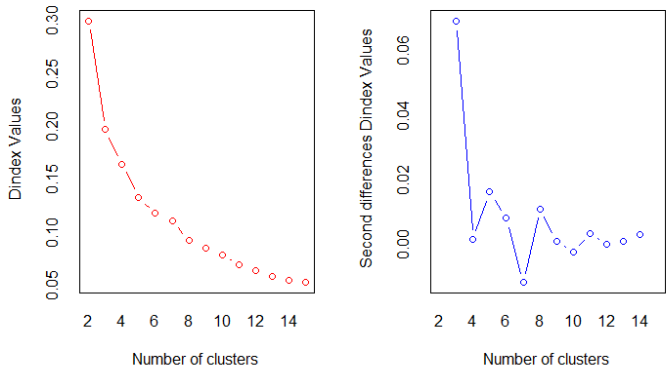

Fig 5.2 Clustering of trips Index

```
Within cluster sum of squares by cluster:
 [1] 125283.06  37479.60  60269.02 155080.58  70306.75 110107.53  87909.39 143937.60  99994.55
[10]  93767.38
 (between_SS / total_SS =  99.2 %)
```

The indexing of the trips helps to understand the distribution of cluster trips and classification based on clustering over the period of time, which gives that the most of the trips were joined in a major cluster to represent that the trips belong to the specific time. Thus, it represents the people usage of trips differs based on their time.

5.4 Distribution of Trips:

The people use the bikes fig 5.4 has revealed a lot of information for the new business model to be prepared. Such that the people mostly using the bikes for commuting purpose in their life as the number of trips were higher based on the count during the weekdays and weekends. Moreover, it's interesting to see that the people are using more bicycle trips during the mid of weekdays. From this insight, we can infer that most of the subscribers would be commuters and they are interested to have weekday trips for office and the weekends trips were to explore or spend time in other cities of the people.

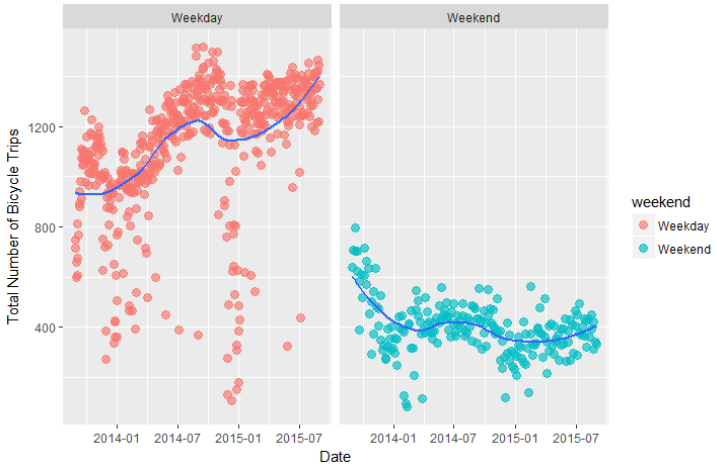

fig 5.4

5.5 Bike Durability mining:

Whenever there is a machinery in a industry, understanding the way of usage fig 5.5 of that helps to aware of business knowledge.. I was curious the understand the that how the bikes were used for the trips and do they need maintenance. The bikes which were used in extensively and also the bikes which are used very less has been figured both the types of cycles would need maintenance. It helps to understand when the *bicycle needs repair* and also to understand the cost spent on bicycles has a profit in the business.

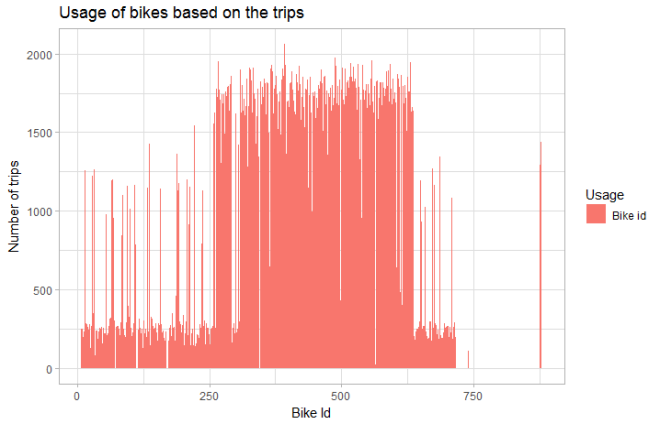

Fig 5.5

As we can see that the bike usage based on the trips, gives information such that the bikes with id 260 to 600 has high usage where other bikes have comparably very less usage. The bikes from id 730 to 870 where relatively used very less and remain useless. The step to improve the usage of bikes will motivate to have a regular maintenance of bikes for the business. It can also be observed from the image that the bikes which are heavily used where the people prefer to have a preference on their bikes and also that the bikes in the San Francisco Bay area need high maintenance where the number of trips are higher than any other city. For Business, the periodical distribution of bike cycles over the period will help to have the bike maintenance cost as low as possible.

6. Conclusion:

Thus the SF Bay area trip dataset has given the wide set of opportunity to understand the pattern of rides and derived a lot of information from mining. The different insights derived from the dataset helps to plan for the future to structure more efficient plans and implement new advanced systems to increase the number of rides and friendly to the people preference[3]. Thus, by the conclusion, the growth of San Francisco bay area bike ride usage has a great potential to reveal new business models which helps a lot of people to go pollution fewer rides and have a healthier life as well.

I would also thank to the kaggle competitors who shared their view of a problem understanding and I also tried their interesting results with my observations, where the algorithm and the methods I proposed **matches perfectly** with other kernel results.

GitRepository:

https://github.com/Vinoth-kumar456/From-Data-to-Insights_DM_M1_Project

References:

1.	https://www.kaggle.com/ievgenii1101/analyzing-duration-of-the-trips	Time-based analysis
2.	https://www.kaggle.com/parryfg/time-based-data-exploration	Duration analysis

** Note: My experimentation results on understanding and applying different kernels are also attached as images.

[3] Note : More Analysis results are shared in images