Airbnb Analytics

Project Report

By

Afreen Patel (011811397)
Prateek Sharma (012416144)
Vinit Singh (012481209)

05/01/2018

CMPE 256
Large Scale Analytics

Guide
Dr. Magdalini Eirinaki

Table of Contents

Chapter 1. Introduction

1.1 Motivation:

The idea is inspired by nature; every creature in this universe having an ability to think, make decisions, and take action does so by learning from past experiences.

Humans use sense organs to get data from the surroundings. We use this data to do a lot of things. One important thing we do knowingly or unknowingly is that we learn from the data and the outcome, which can be used for future references. For example, you know it's not a good idea to miss your friend's birthday, you learn from your past experiences and make sure you don't miss it in the future.

Computers, on the other hand, do not have this ability yet. They can't decide for themselves what's important and what's not. Computers need specific instructions about the steps that need to be exercised. Because of the advent of new computing technologies machine learning today is not like machine learning in the past and has gained a lot of power. This has enabled us to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results – even on a very large scale. All these developments have led to the introduction of a new leg in data science and engineering called Large Scale Analytics.

In this project, we are going to obtain conclusions analyzing and processing large datasets, which could help the company drive its business using its own data. We are choosing Airbnb for this project. Airbnb was founded in 2008, is a global travel community helping people find a place to stay anywhere in the world. It uniquely utilizes technology to economically empower millions of people around to world. They help house owners monetize their spaces, passion and talent to become entrepreneurs. With a global reach in more than 191 countries, it lists apartments and villas to castles, treehouses and B&Bs.

1.2 Objective:

Current Airbnb application supports following features,

1. Become host
2. Book apartment/home

In addition to these functionality, we are suggesting with respect to existing investor, consumer and new investor,

1. Predicting average price of property
2. Suggesting area to invest in properties to get more profit
3. Providing feature for consumer to plan trip according to the variation of price over the season.

1.3 Project Goals:

1.3.1 Where to invest in property to get more profit?

We are adding feature which will help the host to have idea about the property investment. In this we are suggesting the different areas which can be best fit for the investor to make properties available to make more profit, which include what type of property, number of beds etc. These suggestions are based on the different parameter like, property type, number of beds, security deposit, cleaning fees, average price of the property. To achieve this prediction of location we are using different machine learning models varies in supervised category.

1.3.2 What price to have for new property listing?

Average price prediction: Prices of properties on Airbnb vary based on the kind of facilities provided, type of property, the location of the property, and time of booking. Since the platform is extremely competitive, new property owners should be able to list their property at a right price to be in business. If the price is quoted high, chances are that the opening business might be dull, which could demotivate the property owner. At the same time, if the property prices are listed very low, the owners could be missing out on some profits that could be theirs.
We will create a machine learning based solution that will help owners decide an approximate price that they must be quoting for their property. This solution could be used internally by Airbnb to audit and regulate prices of properties already listed too.

1.3.3 When to plan a trip to get the deal in Price?

We have analyzed the Airbnb New York city data from the data set to come up with visual graph and metrics that would help customer to plan his stay at Airbnb so one can get the best deal. Results of the analysis includes graphs and some recommendation for the customer. Analysis results points out what is the maximum average price for given listing and the reason for the observed increase in price i.e. if it was a holiday or a weekend.

Chapter 2. System Design & Implementation Details

2.1 Algorithm selected:

Decision Tree, Linear regression, Random Forest classifier, Extra Trees classifier, Decision Trees classifier, SGD Classifier, Logistic regression

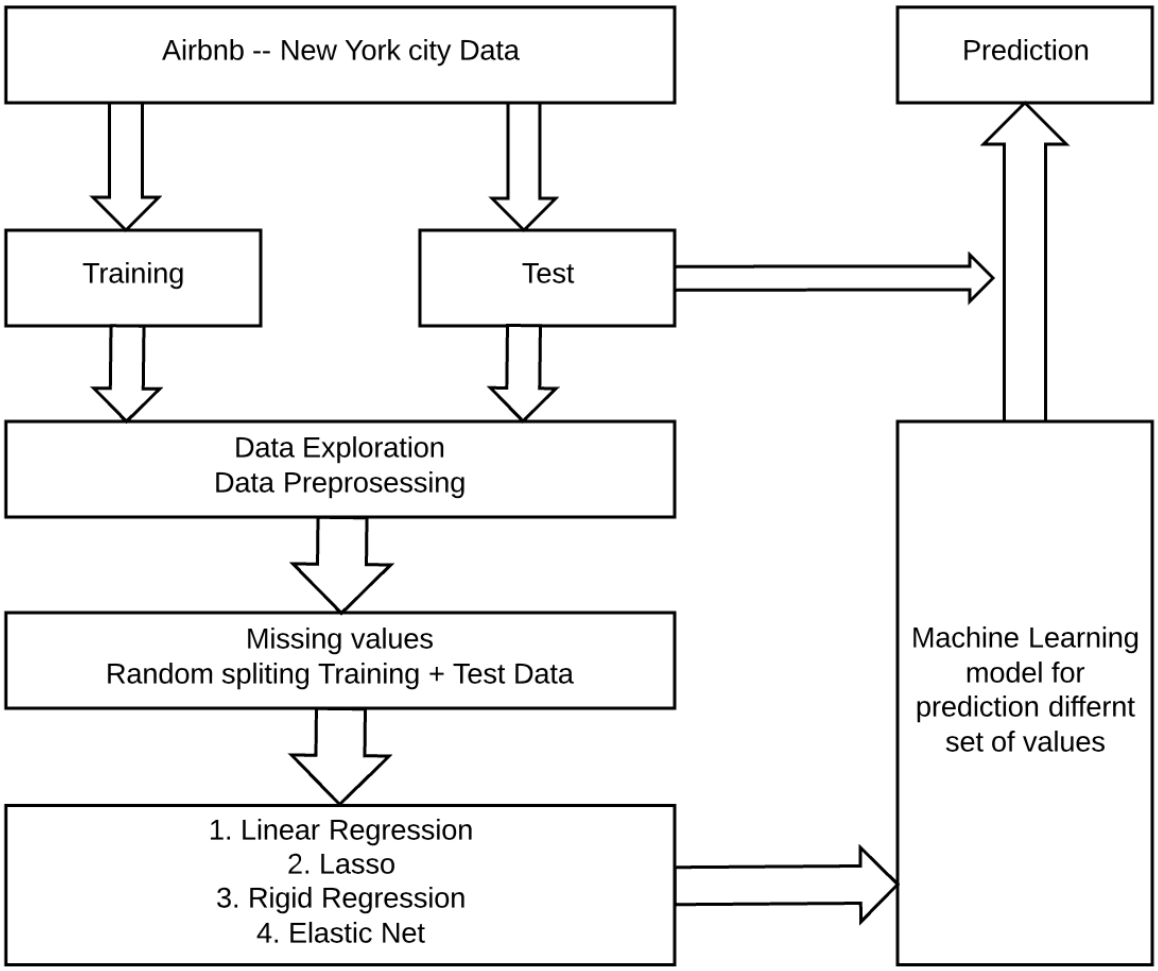
- We started with the dataset exploring, checking the size of the data which has more number of property listings, type of properties, price, booking dates etc.
- We have discussed the about the features of the datasets which can be combined together to generate the valid prediction for the proposed ideas. In this process we have omitted many of the column which were unrelated.
- We tried to figure out relationships between the independent and dependent variables which has the maximum correlation with the help of pearson's correlation matrix.

- At the beginning we started with the simplest approach of Logistic regression to see what type of results we are getting on unprocessed data.
- After the unsatisfactory results we spent time on processing the data like removing the null values, converting categorical values into the dummies feature and removing features which were unnecessary
- After this we have started with the analysis part of the price with the help of the graphical representation using Bokeh libraries.
- We started with the machine learning for predicting the prices for new properties, suggesting area where investor can invest to get maximum profit using sklearn regression model.

2.2 Tool and Technologies Used:

Python, jupyter notebook, sklearn. We have used jupyter notebook in python as team has familiarity with python.

2.3 System Architecture Design



System Architeture

- For the initial step we did the data exploration to get idea about the data and attributes of the complete dataset.
- After this we performed many different type of data cleaning as and when required for predicting. We have removed the columns which has more null values. Replaced null values with 0.

We have applied many different linear regression models to figure out which is the best for prediction

1. Vanilla Linear Regression
2. Ridge Linear Regression
3. Lasso Linear Regression
4. Elastic Net Linear Regression
5. Bayesian ridge Linear Regression
6. Orthogonal Matching Pursuit Linear Regression

Chapter 3. Experiments

Dataset Used: Airbnb Open Data in NYC
Size: Includes 7 csv files (382.29 MB)

Dataset selection:

We have chosen the New York city's dataset from Kaggle for this project. The real estate in New York is very unpredictable and varies around the year. Some of the other factors under consideration were consistent demand, a wide range of apartments, financial capital of the world, effect of the economy directly on the prices, culturally versatile, and a hub for travelers across the world. It's hard to say where the New York market will go in the future, hence prediction of numbers become even more essential.

From an engineering standpoint, the dataset provided by Airbnb is complete and has most of the data public, which is very rare to find in a competitive business like real estate. Airbnb has made sure most of the fields have uniform data (and data types) making it ideal to run our algorithms. Working on clean data can make the model and times efficient and accurate yielding better results. Since the decision-making process directly depends on the quality of the data under processing, a clean data set helps.

3.1 Brief description of Data Set

The data set used in this project includes following CSV file.

- Calendar_detail.csv: Contains details of all Listing in New York along with date price and availability.

- Listings_detail.csv: It includes details of the listing w.r.t summary and description of given Listing.

- Listings_summary.csv: It provides summary of Listings w.r.t its neighbourhood, geographical details, room type, price and availability.

- Neighbourhoods.csv: Provides details of neighbourhoods for a given listing, it included 2 columns neighbourhood_group and neighbourhood. We can get the neighbourhood_group from the listing summary csv for a given Listing and find out Neighbourhood of a given Listing.

- Reviews_detail.csv: Provides reviews from a customer for a given listing id.

- Reviews_summary.csv: Provides dates when individual listing was reviewed by a customer.

3.2 Data preprocessing

The dataset had values that could not be readily used for processing. There were few processing techniques we used to make the data more machine friendly as listed below -

- As a first step, we took columns only that were needed for processing and dropped the rest of the columns from the data frame under consideration.
- Properties that did not have any reviews per month had empty/null values, 0's were added to such properties.
- The same was done for review scores.
- Properties that did not have bedrooms, beds, or price listed were dropped from the data frame.
- The price column had values as a string with a '$' as a prefix. We removed the '$' and converted the rest of the string to float.
- Some columns do not have directly quantifiable values, the machine will have difficulty in categorizing such columns. To solve this problem, we have added dummy columns for all the possible values for that particular field and assigned boolean values according to its presence for a particular property.
- To analysis seasonal variation of price, it was necessary to split the data to get the details of season. To get more information out of the dataset, we split the date for each record as data, month and day. Further the data was grouped to get mean price for each listing over each day of a week and weekends. Used numpy and pandas for cleaning the data set to yield meaningful data for our analysis.

3.3 Methodology

What price to have for new property listing: After the data was cleaned, and the data frame was ready for processing. We investigated on distribution of apartments based on the number of rooms it has. It was found that more than 80 percent of the apartments were hosting one room. To obtain more accuracy, we boiled down our processing to just one room apartments. Similar processing steps can be followed for two, three, four room apartments as well.

The data frame was divided in the ratio of 4:1 as training and test dataset respectively. One of the best things about scikit-learn is that we can easily try a bunch of different linear models on the particular dataset. This will tell how we stand, and what kind of tuning can be done. We will start with six of them: vanilla linear regression, ridge and lasso regressions, Elastic Net, bayesian ridge, and a lesser used one called Orthogonal Matching Pursuit. Results of all the models are compared to see which performs the best.
We have chosen median absolute error, mainly because it makes sense at a glance and is less sensitive to outliers than other metrics.

Analysis of Price for Listing with Seasonal Variation: Analyzed the average price of listings, grouped the data over year and month to evaluate mean price of given listing, used matplotlib to plot the bar charts to visualize average price variation. This helped us understand trend in price over months, next we wanted to extract variation in price for given listing over days of week, and holidays, the algorithm first checked if the date is weekend or holiday and then find average price for every day in a week for a given listing. To visualize the results barplot was used. Last part of analysis included finding average cost of listing during holidays. To do this we used datetime, calendar and holidays libraries. Bar Plot was used visualize the results of analysis.

What are best locality available in New York: The suggestions are based multiple independent attributes of the entire datasets which affects the prices. In this we have performed a scikit-learn, linear regression model to suggest which are the best area in the New York which will return maximum profit. We have tested this with wide range of linear regression model like ridge and lasso regressions, Elastic Net, bayesian ridge, linear estimator, LAR. We compared the models results choose the best model to predict the area and price

Chapter 4. Discussion & Conclusions

4.1 Difficulties faced

- The data set that we selected was huge, due to which large amount of time and effort was utilized in cleaning and exploration of data.
- Due to large number of rows in the dataset and given limited amount of computational capabilities, it caused lot of memory utilization which ultimately lead to time loss during testing and sometimes even system used to run out of memory. To resolve this, we had to process the data step by step and in limited chunks.
- Team faced difficulties to ensure the algorithm worked as expected with such huge data.

- Faced difficulties while setting up the google api for displaying the google maps with the Bokeh library.

4.2 Things that worked well

- Selection of dataset, hugeness of data provided us with option to delete the unwanted data and still have good amount of clean data for processing.
- Preprocessing the data helped team to learn more from the data, it really helped us understand and provide better analysis and predictions for the data.

4.3 Things that didn't work well

- Initially, we planned to sample the data into small chunks, this samples were to be used to train and test the model. But things didn't go as expected, due to the vastness of data the samples that were produced were not able to train our models very well and it involved considerable part of error in prediction that could not be ignored.
- As a team, we had to take to decision to use entire data for training and testing with a ratio of (80% Training and 20% Testing).
- Some algorithm like SVM good amount of time to run and produced results that were not satisfactory, as a team, we had to work a lot to understand the principles on which these algorithms worked which later helped us solving our problems.
- Tried to do the sentimental analysis to predict the area to invest in property. This approach didn't work out as the reviews are not the only factor which will affect the investment decision. We have considered this as regression problem and analyzed over various data attributes.

4.4 Conclusion

- We have explored the different dataset and figure out the relationship between the independent and dependent features. Data exploration also helped us in finding out features which have missing values and categorical attributes which could be converted to binary.
- We learned the techniques to handle the large imbalanced dataset and how to apply regression problems based on it.
- We concluded that depending upon the requirement of the predictions every other model has different output, like in suggestions of the areas LAR regression model is giving the best median value. In other hand for price prediction for new property listing, Bayesridge regression model has return a best absolute value with maximum accuracy.
- For the seasonal price analysis, we narrowed down to the best results with the support vector classification.

Chapter 5. Project Plan/ Task Distribution

Airbnb Analytics	
Task	**Responsibility**
Dataset selection	All
Data Exploration	All
Data Preprocessing	All
Research on Classification and Regression algorithms	All
Scikit Learn regression models	Afreen, Prateek
Linear Regression	Afreen, Vinit
Random Forest Classifier	Vinit, Afreen
Decision Tree Classifier	Prateek, Vinit
Linear Regressions models	Afreen, Prateek
Support vector classifier	Vinit, Prateek
Documentation	All

Chapter 6. References

1. https://www.kaggle.com/peterzhou/airbnb-open-data-in-nyc/data
2. http://scikit-learn.org/stable/modules/model_evaluation.html
3. http://scikit-learn.org/stable/modules/grid_search.html#alternatives-to-brute-force-parameter-search
4. http://scikit-learn.org/stable/modules/linear_model.html#ridge-regression
5. http://scikit-learn.org/stable/modules/linear_model.html#lasso
6. http://scikit-learn.org/stable/modules/linear_model.html#lars-lasso
7. http://scikit-learn.org/stable/modules/linear_model.html#elastic-net
8. http://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html
9. http://scikit-learn.org/stable/modules/svm.html
10. http://scikit-learn.org/stable/tutorial/basic/tutorial.html
11. https://matplotlib.org/
12. http://ipython.readthedocs.io/en/stable/interactive/plotting.html
13. https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python
14. https://matplotlib.org/api/_as_gen/matplotlib.pyplot.bar.html