ITCS5156 - Applied Machine Learning Project Proposal

Prasanna Shanmugakumarasamy
pshanmu4@uncc.edu

Project Title and Sources

- Title: **Predicting Airbnb Listing Price Across New York**
- Research Paper Source:
  - Journal Name: **Researchgate.net**
  - Publication Year: **2020**
  - Author: **Brahmaih, Kala**
- UNCC Student doing this project:
  - **Prasanna Shanmugakumarasamy**

Problem and Dataset

Problem:

Pricing a rental property on Airbnb is a difficult task for the owner because the number of customers on the premises is determined. On the other hand, customers must assess the offered price with minimal knowledge of the property's optimum value. In order to help the landlord and consumers with price estimation, **this paper and my project seeks to build a robust pricing model** utilizing computer intelligence, deep learning and natural language processing techniques, despite the minimum accessible description of the house.

Datasets:

https://www.kaggle.com/duygut/airbnb-nyc-price-prediction/data

http://insideairbnb.com/get-the-data.html (Scroll down to New York)

Motivation

Airbnb is a worldwide rental service provider. However, every city has its own dynamics and specialties that makes the city's properties pricing strategies and type of listings very unique. New York city is one of the unique metropolitan cities in USA that has tourists, residents, floating populations etc., amidst a fast-paced environment. It is very interesting to study the different types of properties around, how they're priced and what story the projected trend tells us for the various locations of the city and the various properties.

On the other hand, Airbnb has a very organized and wide range of dataset starting from the listings, the availability by calendar days and customer reviews. The scope of analysis is very wide as well.

Lastly, the paper i have chosen introduces me to several machine learning techniques and toolsets that I've not explored yet.

Together all these reasons stated above, it motivated me to pick this topic and explore further on this project..

Related Works

Primary Work (to Duplicate):

- Brahmaih, Kala. "Predicting Airbnb Listing Price Across New York". *Researchgate.net*. (2020).

Reference Papers:

- Rezazadeh Kalehbasti, Pouya et al. "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis". *Machine Learning and Knowledge Extraction*. (2021): 173–184.

  Why? *This paper discusses sentiment analysis in addition to the Price Prediction using Machine learning techniques. Text Mining and Analysis is an area that is covered additionally. It uses the customer reviews dataset to deduce this information. Also, the reviews dataset can be condensed and combined with the listings dataset.*

- duygut. “Airbnb NYC Price Prediction | Kaggle.” *Kaggle: Your Machine Learning and Data Science Community*, Kaggle, 22 Oct. 2019, https://www.kaggle.com/duygut/airbnb-nyc-price-prediction.

  Why? *This is one of the Kaggle Work that is most discussed. This uses Ridge, Lasso and Elasticnet regression models that were discussed in the class. It will be a good place to apply and evaluate. Also, it uses GridSearchCV algorithm that helps identify best hyperparameters for each model.*

- Dan Wang and Juan L Nicolau. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com. International Journal of Hospitality Management, 62:120–131, 2017

  Why? *This is one of the most referred research paper amongs every research project that uses Airbnb dataset. It covers a wide range of cities and since each city is different the exploratory analysis are entirely different and the path to modeing and analysis are different as well. This serves as a base study and apply some techniques based on the results for the New York city that we’re considering to analyze.*

Methods of the Selected Paper

Dataset Study and Exploratory Data Analysis:

The paper describes the different attributes of the datatset and its nature and applicability of them to the New York city. Some of the attributes such as Amenities type, property type, Room availabilitiy and neighborhood districts are categorical variables that are used in various Exploratory data analysis steps. A map of the neighborhood is also developed to understand Geography of the airbnb listings.

Data Cleaning and Preparation:

Variables that are redundant and duplicates are removed. Few attributes are transformed to suppress skewness that can introduce bias in the data.

Methods of the Selected Paper

..Contd

Modeling Techniques:

Below Modeling techniques were used to predict listing prices.

- ***Linear Regression:*** Estimation of values of the data representation coefficients.
- ***Gradient Boosted Regression:*** This method groups all the weak predictive models and built the accurate model,which typically are decision trees. This combines both the decision tree and the boosting algorithms by considering the previously built decisions trees
- ***XGBoost:*** XGBoost is a decision tree assembly boosttree algorithm.It varies in two ways from the random trees.Random forests first construct a tree on their own while XGBoost produces a tree on a single basis

Results and Prediction:

R square and RMSE values were used to assess the accuracy of the models and to compare performances of the models.

Methods of the Selected Paper

..Contd

Why it makes sense?

Price is a continuous attribute and is not a binary or a multi value categorical attribute. The continuous attribute cases requires a regression method or a decision tree based models to predict the trend. These techniques also help understand the correlation between variables and the importance of variables from their coefficients.

So, the discussed methods convincingly can predict prices and the dependent variables that influences the price.

Methods of Reproduction

How will I reproduce the work ?

- Study the Data Dictionary and understand the Dataset better.
- Look into the research paper for the Exploratory Data Analysis and Data Preparation.
- Review how i can duplicate or find alternative for the apparatus used such as Azure Databricks replacement or use of PySpark to read or transform dataset. I have to find ways to do them directly in Jupyter Notebook with Pandas.
- Understand the different Regression techniques and read about the new techniques that were not discussed in the class such as, XGBoost.
- Study and Duplicate the Price Prediction based on the model developed.
- Explanation of the models and arrive at the conclusion.

Timeline

7-Feb	Complete Dataset and Data Dictionary study and complete Exploratory Data Analysis.
14-Feb	Data Preparation for Modeling and attempt a few Models
21-Feb	Complete Modeling and Performance Review of Models
28-Feb	Explain the models, Conclude
02-Mar	Complete Project Report and Submit

Expectations

- Higher level of Programming with Python and handling Jupyter Notebook
- Conversant with Exploratory Data Analysis and Data Visualization modules
- Learn the Regression techniques such as Gradient Boosted Regression, XGBoost and Linear Regression Variants (stepwise).
- Improved knowledge on Training and Testing models.
- Explore GeoPandas and iPlot libraries.
- Learn How to evaluate and explain models.
- Overall execution of a Machine Learning Project cycle.