# Recommended Categorical Price and Availability Prediction using Deep Learning Neural Networks in Airbnb Toronto Listings Dataset

Saad Hasan, Ioannis Lambadaris
*Department of Systems and Computer Engineering,*
*Carleton University,*
*K1S5B6,Ottawa,ON,Canada,*
*Email:saadhasan3@cmail.carleton.ca,*
*ioannis@sce.carleton.ca*

Omair Shafiq
*School of Information & Technology ,*
*Carleton University,*
*K1S5B6,Ottawa,ON,Canada,*
*Email:omairshafiq@cunet.carleton.ca*

***Abstract*—This research work is focusing on the idea of implementing machine learning algorithms in the field of business as these days Business intelligence is one of the top growing field as of today. Airbnb is an online marketplace and top firm for renting house or apartment which is one of the most widely used specially by the tourist and travelers. We made a prototype of our model which is based on deep learning neural network to predict the Price classes of affordable , expensive and cheap on the certain features. In the second part of the project we applied the same model to predict the availability and do the comparison of other classification algorithms like Random forest and XGBoost. With the help of accuracy our model works and gave better result as compared to other algorithms. In the existed work highly focused more on visualization and predicting the original price value by considering the regression problem and achieved very poor accuracy but transformation of problem from regression to classification improved the accuracy significantly and made our model more efficient as compared to previous work. The main ideology behind the transformation to classification is that people have specific range of price in their mind to rent a house or an apartment through Airbnb and that feature is missing in Airbnb application. Implementing this feature will increase the profit for the Airbnb company as it will increase the customer experience.**

## I. INTRODUCTION

In the year 2008, which is the birth year of the firm known as Airbnb which is widely used for home staying an on
line marketplace and a service of hospitality service. Most of the visitors are usually travelling inter cities and International tourism gained the popularity which is an important factor in the economy of the country as well. Sharing economy, novelty , home benefits, local authenticity , privacy and prices are the most important factors in the Airbnb which attracts tourists and changed their mind to use Airbnb instead of other motel and local services [1].

For the last couple of years, its popularity plummeted to 33.9 million of users alone in the United States and further increased to 45.6 million by 2022 in USA which made the service popular among the tourists [2]. In the year, 2017 it surpassed 31 billion dollars and around 4.5 million listings located in 191 countries [3].

Using data on previous Airbnb listings, optimal price prediction which is widely used in the past by the different researchers mentioned in the literature review section Canada will be performed. The user can use our initial price to start with and then make necessary changes based on interaction with the customers to reach a price that is optimal in terms of the hosts profitability and still affordable to the guests. From a real estate investors point of view, maximum revenue generation can be studied by understanding how the factors like amenities, location and availability can influence a price class. Furthermore, analysis will be performed to determine the likelihood of a listings availability for potential guests to consider while making a booking. The models provided by this project can be helpful to the internal pricing tools that Airbnb provides to its hosts.

Implementation of Price and availability as a new feature which will helps the following benefits for the business which include

**Revenue:** The most important thing is the revenue which is the main goal of the business and developing new feature will impact the revenue of the business.

**Customer Satisfaction Need:** The most important thing in the business is the customer satisfaction which will increase the profit in the business.

**Customer Price range:** As most of the people focusing on the price range rather than the specific price. Recommendation about price range also help the customer to understand the budget and his pocket to get the room.

Whether the host who is hosting a room or a house or condo or a tourist who is renting the house or a room , both must be satisfied to increase the popularity and profit of the business.

Airbnb also used by property owner which lease their apartment for short terms to get more benefit in less days.

**Overview of Proposed Solution:** With the help of predicting classes, we can estimate and understand the factors which are contributing the most impact on Prices and availability of using Airbnb whether bungalow, house or a room by applying the deep learning neural network as using this model will also helpful to learn the non-linearity in the dataset as most of the data in the real world is nonlinear. With the help of finding the factors our model can only give the benefit to the customers

who rent a house or an apartment in Airbnb or the host who are giving their apartment or room for Airbnb but also for the Airbnb firm as the new feature of classification based on price range is might handy for the customers. This thing will increase the customers and hosts as new people will come to the Airbnb will increase the business too.

## II. LITERATURE REVIEW

Airbnb is an American privately-owned firm working as a middle ware to book a room or services from website or any android or IOS application. The good thing with this application is that people can easily book their room when they are travelling from one place to another. It could be the people who are migrating from one city to another or it could be the people who are coming from one country to another. For example, a person who has an interview in Toronto and he is living in Ottawa. So, he needs to book a room for the visit. In the past, people used to go through each website to see the availability and price of the room. But these days due to the Airbnb and other platforms like VRBO and HomeAway helps the public to get their desired room with their desired prices and availability of the room and saving time instead of visiting and go through each website to see what is affordable or not.

Several studies worked on this domain to improve the system which will be beneficial not only for the Airbnb firm but also for the public.

Ki-Hong Choi1[4] et al. uses panel regression model to see the what would be the effect of Airbnb listing on the Korea 's economy. They selected the top 3 cities which include Seoul, Busan and Jeju. They focused on the idea that if the hotel revenue increases then it means the tourists increased which will be beneficial for the economy of the Korea and helped them to maintain the economy.

The main strength in this paper [4] is utilizing not only Airbnb data set but also taking data from the Tourism Knowledge and Information System and from economic statistics systems and Seoul Money Brokerage Services which is helpful to the Macro Economic Statistics with focusing four main cities to see the full picture in Korea instead of focusing on one cities [4].

The main weakness in this paper is that instead of utilizing other algorithms they just used multiple linear regression for the analysis but there are other algorithms could be used to get the better result. [4]

Laurent Son Nguyen et al. [5] used Airbnb images to predict imitations in Airbnb dataset but only focused on the dataset images found in Switzerland and Mexico. They [5] also performed clustering analysis based on the features which are physical and emotional features which is automating the intuition of human with the help of images collected from Airbnb dataset which is focused only in Switzerland and Mexico.

The main advantage in this paper [5] is the technique they used is convolutional neural network which is very efficient and good method to use for images. The main weakness and the problem in this paper [5] although they used very complex and robust method for the images application, but the accuracy is too low in the form of R squared which is 0.34 overall and it is too less for the good result.

Longhua Guo et al. [6] used rating and records for the room hosting to see the effect of having many users shared the room and what would be the effect on the economy. They [6] utilized the statistical concepts confidence, normalized variance and other statically techniques to see the effectiveness of their proposed method. The two algorithms they proposed based on weighted matrix factorization and based on similarity with other conditions to check the accuracy on the real data as the problem identified as NP hard problem.

The strength of this paper [6]is that they target the problem of NP hard which is in the form of matching two partners sharing the room and, they proposed two matching algorithms which are greedy algorithms but never proposed by anyone before. The weakness in this paper is that they only proposed the solution which haven't tested by anyone before.

Qian Zhou et al. [7] did analysis on the profiles of the customers using Airbnb globally and also analyzed the pattern and identified the users from different countries with the help of machine learning algorithms like Xgboost, random forest , c4.5 and Bayes net which will be helpful for sharing economy concept from customer view by considering around 43.8 million of Airbnb users profiles.

For the very first time, they used [7] an idea of checking the profiles of people in Airbnb. Also, more than one algorithm they used like random forest , xgboost, c4.5 and Bayes net for the prediction and for the comparison of result. The weakness in this paper [7] is that although they used profiles from mall over the world, but they missed the analysis of profiles behavior online and more complex algorithms like deep learning which are suitable for this application is missing.

Moloud Abdar et al. [8] uses the idea of sharing economy in the Airbnb dataset and with the help of statistical techniques and applied the rating matching method to the rating and location and found that countries USA, UK, Germany, Australia and Russia are the important places for the 5 star hotel housing.

They [8] considered around 10 countries and made the model based on the result which is the combination of 10 countries listing s from Airbnb which is challenging and difficult to analyze. Also, they used regional characteristics with the help of location to see the human behavior and, they used economy sharing business principle which is a new trend these days. The weak point in this paper is that based on combining data and different rates followed in different countries and preferences. Their model will not be helpful in all 10 countries because of the nature of the data in each country is different.

Czesław Adamiak et al. [9] focused on the regression analysis of Airbnb listing in Spain and with the help of spatial resolution they found out the pattern in the listing of the Airbnb in Spain. They focused on the location-based system which will show the locations where investment is worthy for tourism.

They [9] achieved very high R squared combined for the model and, they focused on novel approach of spatial resolution in Spain which is make sense as they don't focus globally but used one country instead as one country have kind of similar characteristics. The weakest point in their [9] research is that they don't focus on the price and availability. Furthermore, they used only Linear regression model and do not use other algorithms to compare and better result.

Giovanni Quattrone et al. [10] primarily focused on the prediction of spatial penetration in the listing of Airbnb by

choosing the eight most populated cities in the United States based on the geographic location. They also classified the talented and creative classes based on location in the form of Quantitative Analysis.
The main advantage in this [10] paper is that they analyzed the major and important cities in the United States for spatial Analysis. They also worked on the new concept of spatial analysis like [10]. But focused in the U.S and achieved very meaningful results and analyzed the relation between geography, social and economic relation in the Airbnb in the important cities in the U.S. The main drawback in this paper is that the work is existed and, they only used the Multiple linear regression and did not focus on other good algorithms [10].
Hanna Lee et al. [11] find out the impact in the sharing economy of Airbnb hosts with the help of psychologically. Structural Equation Modeling and confirmatory factor model utilized for validation while one tailed method of hypothesis test used for finding the attachment of Airbnb hosts.
The new concept of psychology is used in the Airbnb business idea of economy sharing, which is the strongest point in the research and, they [11] got the data from questionnaire instead of taking exiting data available online which gives uniqueness to their method. Furthermore, the problem in their [11] research is that they missed the observations because the information form questionnaire is highly dependent on the persons response and response could be false and achieved very low R squared which is essential to work for the model well.

Wang Min et al. [12] uses Airbnb details of the data and tried to use Extended theory of acceptance and hypothesis testing for finding the factors contributing the customers to use Airbnb.
They [12] used hypothesis testing in Airbnb dataset from the text and their main focused is on the consumer relationship to select Airbnb or not which is very strong point with respect to business perspective. The weakness in this paper [12] they focused only on two personality factors and missing other personality factors which contribute the factors to select Airbnb by the consumers. Another weakness in this paper is that they neglect probability sampled data so only small proportion of the dataset is used.
Raza Hasan et al. [13] also focused on the idea of economy sharing concept and with the help of interviews conducted they successfully found out the factors contributing the success in the economy sharing.
The business model they [13] developed is the part of the concept economy sharing which has not only application in renting a house or room in Airbnb or ride share in Uber when the customers share the same ride. But the defect in their research [13] is that they used empirical data instead of real data to understand how much the model is good for the real-world data.
Kun Wang, et al [14] uses different machine learning classification algorithms including naïve Bayes, support vector and machines and decision trees in the smart grid applications to predict the price with very promising result come from differential evaluation-based support vector machines as compared to other models.
The strongest thing in this [14]paper is they worked in the topic of price forecasting in the smart grid which is a hot topic and very useful. Also, they [14] used not just an ordinary support vector Machines but do some modification in the algorithm which gives better results in the accuracy and precision as compared to ordinary support vector machines and other algorithms. The problem in this paper although they achieved very good result because of 50,000 data points. If the data points are less, then their technique will not give very good accuracy.

Subhajit Sidhanta et al. [15] uses optimization in the framework of Apache Spark to reduce the cloud resources and improves the calculation time in many applications in the cloud architecture including Hadoop Distributed File systems.
The key point in this [15] paper is that with their model is optimizing the process and calculate the time of completion and independent of the size and type and complexity of the data. Also, the category of jobs is also clustered with 98 percent accuracy by dividing the jobs based on similarity.
Malav Shastri et al [16] proposed and analyzed the stock exchange data with the help HIVE which is a framework of Hadoop Map reduce and predict the prices of market with the help of Artificial Neural Networks.
The most important key points in this paper [16] is that they used textual analysis to predict the type of the sentence whether positive or negative and get 98 percent accurate result. Also, they predict the prices of stock market with the help of ANN and achieved with around 91 percent accurate result. Although with such accuracy there are still lot factors which are missing and have strong relationship prices of stock should be considered in textual analysis.
Ying Yu et al. [17] predict the economy based on seasonal auto regressive integrated moving averages and Neuron model with dendritic nonlinearity for forecasting. They [17] developed which is the mixture of these algorithm and did time series forecasting in Tourism Industry.
The most challenging and key point is that although the relation in data is not linear so that's why they used the combination of SARIMA and DNN model to handle nonlinear correlation and the model handles both whether the linear and nonlinear datasets [17]. The main weakness is that in the paper [17] they did not discuss about the accuracy in the form of R squared for forecasting economy trend in tourism industry.
Mathias Longo et al. [18] uses Long Short-Term Memory for the prediction of availability of energy and developed a model which is forecasting hourly Prediction. The model is based on the recurrent artificial neural network because of the data is highly dependent on time and in each timeline is dependent [18].
Strength in this paper [18]is that they utilized Long short-term memory recurrent neural network to predict the availability which is one of the good algorithms for time series and time dependent forecasting. But still they ignore the information which could be taken from Time series Analysis.
Sheikh Mohammad Idrees et al. [19] uses time series Auto regressive integrated moving average model which is an effective method for time series forecasting and applied it to predict the future prices in the stock market focusing mainly in India.
The strength in this paper [19] is that they used the data with high P-value which is around 0.90 and 0.86 for both stock data of Nifty and Sensex respectively. They achieved around 5 X squared for both stock data which states even with such

uncertainty their model has no problem of auto correlation. But still with that model they [19] need to clean more data to improve the model and their model work focus only on two stock market, not all the stock market in India.

Mohsin Munir et al. utilized [20] deep learning anomaly-based approach in the multiple applications of internet of thing for the time series prediction. They [20] almost used 15 algorithms for the evaluation and 10 benchmarks of anomaly detection from these algorithms evaluate. The deep learning method include the convolutional neural network to predict the time series forecasting and these predictions is helpful in next module of anomaly detection part to diagnose the abnormal time stamps in the data [20].

They [20] checked on different data set on twitter, yahoo and other 15 datasets. Their novel method has tendency to overcome the issue of minor data contamination which is usually and covering many domains including different datasets like health care, internet traffic, online advertisement and other etc. The main weakness in this paper is that there model is not adapt for time series analysis of anomaly detection. Also , they [20] missed many preprocessing techniques to make the model better.

Tongtong Yuan et al. [21] uses adaptive hashing to solve the problem of high dimensional hyperspace to low dimensional space using principal component analysis as the preserved hashing is an NP hard problem. They [21] proposed the method unique method to handle the high dimensionality along with handling unbalanced data for an image.

The most important point in this paper [21] as this paper targets the NP hard problem, which is only few people work on that also, they used adaptive hashing method which is novel approach. The weak point in their research [21] is that although with the new idea , they achieved maximum of 66.7 percent precision which is still less for the model.

Neha Bharill et al. [22] work on Apache Spark which is an important cloud resource working based on cloud computing framework and widely used in IT industry. They [22] used Scalable random sampling with iterative optimized fuzzy c means algorithm to solve the problem of large data clustering on apache spark clusters.

Instead of just applying their algorithm on one data set example. They [22] tried on different including handwritten images, skin data set and others which have more than 1 billion of instances which has more than 25 gig bytes of data. Also, for clustering , the time is reduced dramatically by removing the membership matrix. Finding the number of clusters is always questioned in the data science and for big data it considers more difficult based on quality and quantity. For this paper [22] an optimized measure for the clusters to validate, are missing.

Ellie Ordway et al. [23] mainly focus on performance of the auto encoder which will help in the significant reduction of heuristic search. They [23] focused on unsupervised in the form of anomaly detection as in the past, no work done on unsupervised only supervised learning is appeared in the research.

The strength of this paper [23] is that their focused on the unsupervised clustering works very well and successful in detecting botnets, brute force and other type of cyber-attacks and challenging in the domain of cyber security. Auto Encoder and hyper parameter tuning is an essential for the complex data intensive but it also time consuming and more memory is required.

## III. COMPARATIVE ANALYSIS

| No. | Year | Type | Method | Data Set Details | Finding & Result |
|---|---|---|---|---|---|
| [4] | 2015 | Airbnb and hotel revue relationship | Panel Regression model | Tourism Knowledge & Information System, Airbnb Economic, Statistics System2 and Seoul Money Brokerage Services | They achieved R squared in different cities as effect of Airbnb listing in revenue of the hotels and combining R squared in Korea as 0.448, 0.442, 0.229, 0.294, 0.298 and 0.145 in the categories of whole, Luxury, Upscale, midscale, Economy and budget respectively. |
| [5]. | 2018 | Clustering using Airbnb photos | Deep learning Convolutional Neural Network | Images retrieved from Airbnb Images | Clustering analysis of Airbnb photos in the region of Mexico and Switzerland which include the three clusters and achieved very low accuracy of predicting ambiance and other attributes. |
| [6]. | 2018 | Matching scheme for Airbnb users | Greedy Algorithms and customized algorithms | Randomly generated sample dataset. | Algorithms they proposed based on weighted matrix factorization along with the help of optimization is very useful on the real data. If utility of user is considered the model is more accurate and Social Influence caused a more significant improvement and beneficial for the platform. |
| [7]. | 2017 | Airbnb and sharing economy from the customers view | XGBoost, Random Forest, C4.5 and Bayes net | Airbnb user profiles been crawled from Airbnb | Instead of taking tweets in a separate country. They collected all tweets globally in Airbnb and achieved precision of 0.79,0.74,0.78, 0.77 in Xgboost, Random Forest, C4.5 and Bayes net respectively to predict the people who are suing Airbnb worldwide. |
| [8]. | 2017 | Rating of accommodation in Airbnb | Rating Matching Rate | Airbnb dataset and reviews of customers | Crowd Preference Mining Model with the help of Rating Matching Rate concept on Airbnb data on important features like Prototype , room type and rating and what categories are more important on human behavior. |
| [9]. | 2018 | Spatial analysis in Airbnb listing in Spain | Multiple Linear Regression | Airbnb dataset listing in Spain with the help of web scraping | Successfully implemented a regression model using multiple linear regression to the Airbnb listing in Spain by finding the correlation factors in the data with respect to location and find the adjusted R squared of 0.92 in total listing |
| [10]. | 2018 | Spatial analysis in Airbnb Listing in USA | Multiple Linear Regression | Airbnb dataset in U.S | Achieved accuracy of 72.5 percent in the spatial analysis in the listing of Airbnb in the major cities of U.S by applying multiple linear regression and by creating the classes based on availability of Airbnb in the certain areas of cities. |
| [11]. | 2019 | Airbnb Host impact on sharing economy | Hypothesis Testing | Online research Data | By using the online research question, found the detailed impact of people hosting Airbnb and their attachment to the Airbnb and factors and their contribution by developing a model based on hypothesis testing and achieved R squared of 0.762 in psychological ownership |
| [12]. | 2018 | Consumer selection of Airbnb | Unified Theory of acceptance & Hypothesis Testing | Textual comments of Airbnb customers | By considering the concept of economy sharing , they achieved to model based on customer intended to use Airbnb and tested their model and achieved 0.70 R squared and factors contributing are Performance, motivation and price value. |
| [13]. | 2016 | Economy Sharing criteria | Detailed Analysis | Data gathered with the help of interviews from Uber drivers and Airbnb customers | By using the same concept of economy sharing , they found out the numerous factors to increase the success rate in the domain which include price, environment , digital platforms are the most important factors. |

In most of the existed research work, people work and utilizing the dataset in the U.S, Japan and Spain cities and focused more on visualization, spatial analysis , economy sharing concepts and price prediction in the regression analysis but with poor accuracy which is around 70 percent accuracy.

## IV. RESEARCH GAP

In most of the literature review people focused on sharing economy concept applied in Airbnb data sets. Others used the Images to perform clustering based on location. Most important thing in the dataset of Airbnb is the price categories and availability which is not found in most of the research.
The most important in the data analysis part is finding the meaning result from the raw data. To understand the data, most of the people used different machine learning algorithms like XGBoost, multiple linear regression, Convolutional Neural Network and other algorithms has been used but most of them are used and focused on spatial resolution. As the data provide by the Airbnb is hug gigantic and enormous in terms of dimension and sizes. The prediction of price using regression in the literature and on the web has already done on different cities like Seattle, but their prediction is very low like less than 70 percent accuracy. So, accuracy must be improved to add the new feature in the application.

## V. PROBLEM STATEMENT

In most of the cases price and availability is always missed in the literature and previous work. Let's analyzed the data available on Airbnb and its features. The reason for choosing is Toronto as this city has the biggest data points and Ottawa has around 3000 data points which are very less to use deep earning technique like Artificial Neural Network. So, we only considered Toronto. So, we transformed
Initially there are 20303 data points and has 106 columns. As most of the features are either have so many missing values more than 50 percent or repeated or the column with more than 80 percent same categorical values has been removed from the data using R script.
We finally reduced to 20303 rows and 28 features which are going to be further analyzed and removed in the feature extraction techniques used in algorithms.

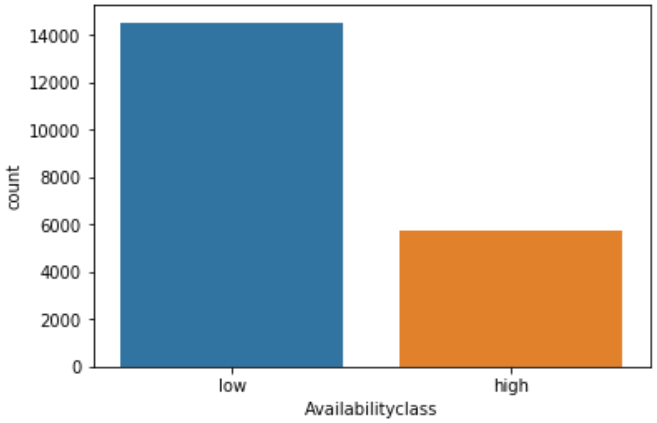


Figure.1: Showing the dependent class availability in the dataset

Figure.1 is showing the how many numbers of these two classes appeared in the dataset. So, with the help of visualization above we can say that around 14000 times the class low appeared in the availability while around 6000 times the class high availability appears. So, most of the time class low appears.

Similarly, there is another class in the dataset which is Price category. It has three categories ; one is good which is also the cheaper class and appeared the most in the dataset. Similarly, the second class is affordable which category appear in count after good class which represents the range of prices which has medium in range. The third class is expensive which represents the prices which are expensive for the residents in the figure. 2

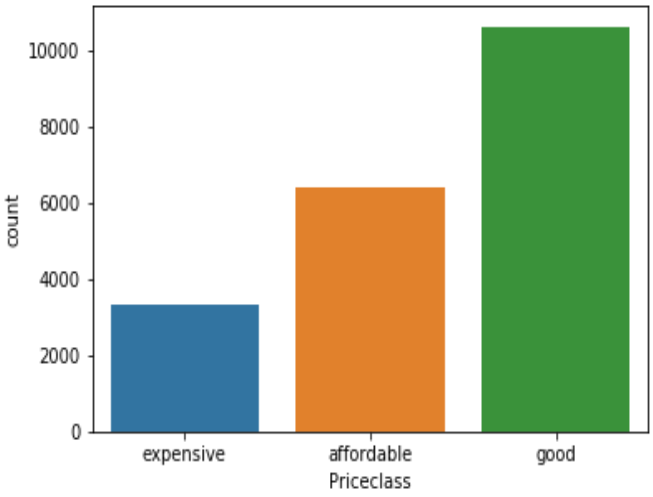


Figure.2: Showing the Price classes categories

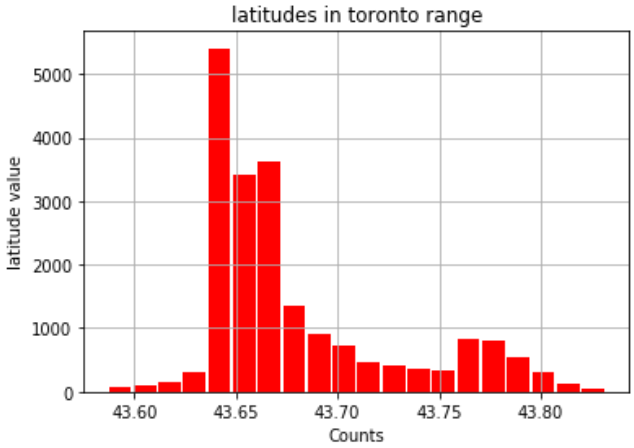


Figure:3 Showing the latitudes range in Toronto

The figure.3 is showing the ranges of the latitudes covered in the location of Toronto. So, most of the locations has latitudes from 43.60 to 43.80 which is the exact location of the Toronto and appeared in the data set.

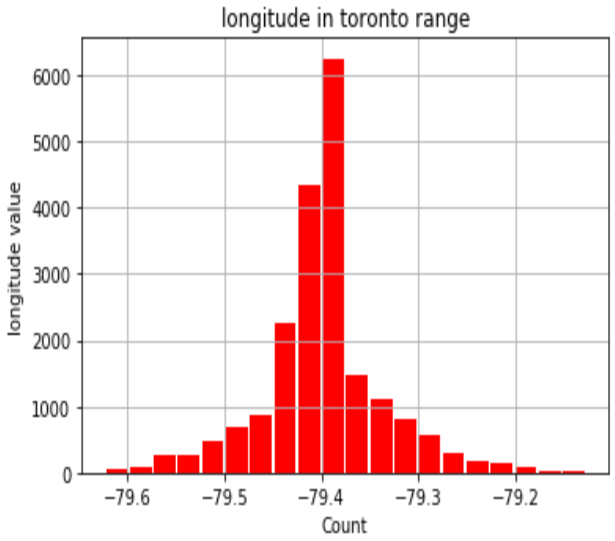


Figure.4: Showing the longitudes in Toronto

Figure.4 is showing the longitude ranges in Toronto appeared in the dataset as it is like the normal distribution.

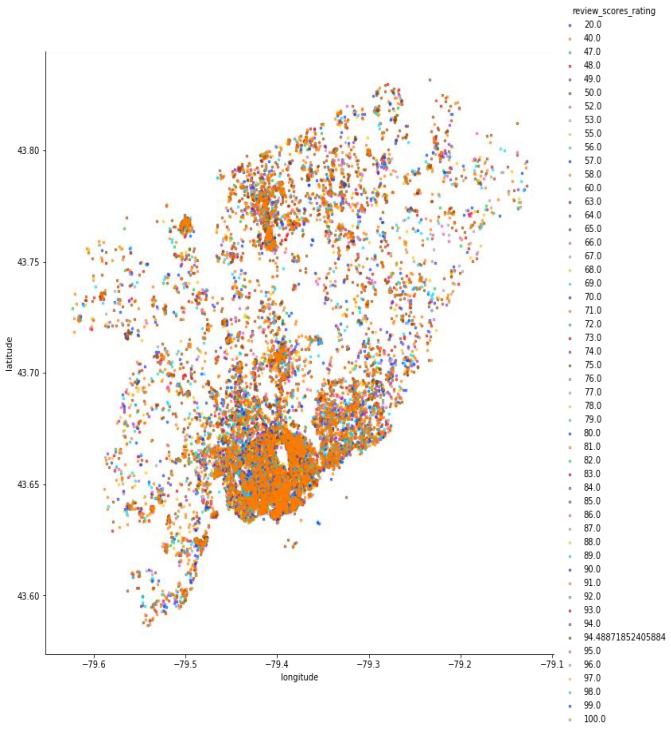


Figure.5: Appearance of coordinates vs review rating

Figure.5 showing the latitude and lognitude corordiantes vs the rating of the reviews about the location. If we take a closer look in the graph which is showing that around -79.4 longitude and -49.70 has 100 percent rating appeared in orange colour.

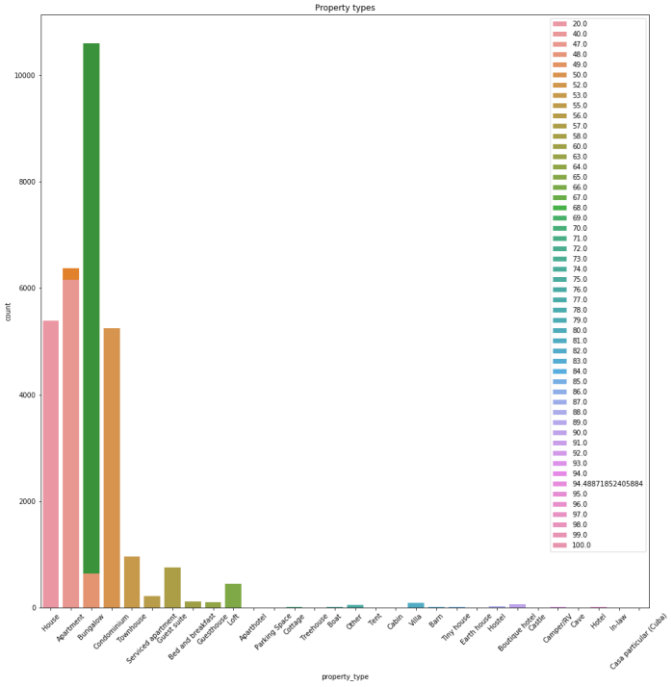


Figure.6: Showing the types of property of Airbnb vs rating

The top four categories of property type appeared in the dataset are House, Apartment , Bungalow and condominium. Bungalow appeared the most, but it doesn't have good rating as green color showing the region of around 65 – 70 percent of rating score. House and Apartment has good rating which is 100 percent. Similarly, the Condo appeared in the rating with respect to the review

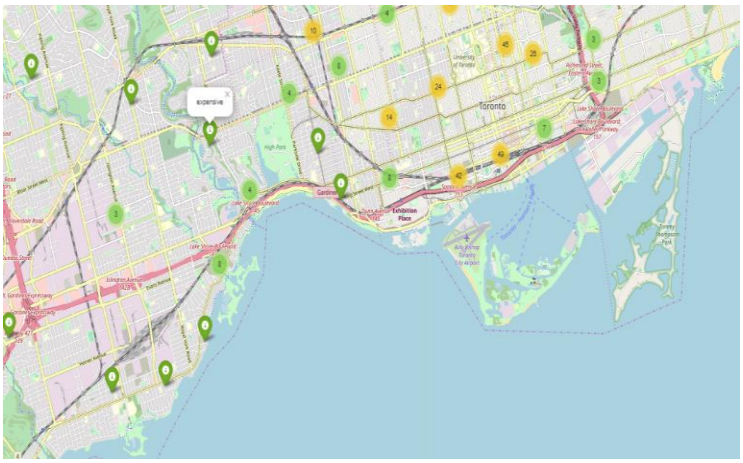


Figure.7: Showing the Price class in the Toronto Area

With the help of installing libraries of folium and base map along with the lib-geos for installing in the directory to map the coordinates of Toronto by using latitude and longitude and plot with respect to Price class. We plot 500 values from last rows to show the exact location of the property and its significance whether it is expensive , good or affordable. Clicking on the location based on latitude and longitude show you the exact type of the class that property appears.

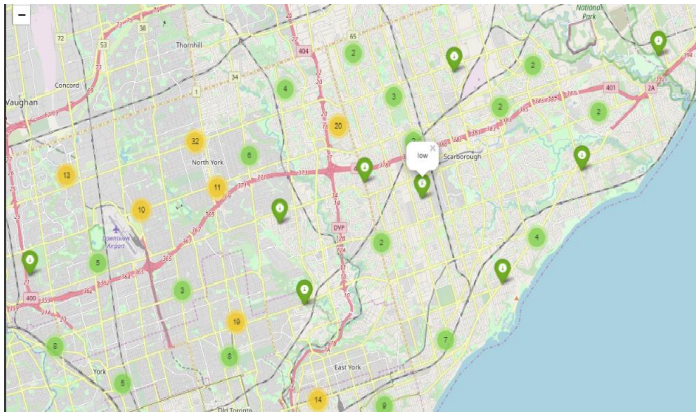

Figure.8: Showing the Availability class in Toronto Area

Similarly, the figure.8 is also showing the same Toronto Area by using the same libraries used in the figure.7 but this time we are mapping with respect to the availability in the area of Toronto with respect to the certain location based on the latitude and longitude of the city.

In our first part of our work which requires the literature survey and find the gap, So based on analysis and comparing different existed work earlier and with the help of Data Visualization , we can say that Price and availability are the two major categories for the Person to find the room in Airbnb. People usually have a range of Prices which they can afford based on their Price range and availability range.

The second phase of the phase is to clean the data which requires replacement through mean , omitting missing values and removing duplicate values in rows and columns. Unmeaning columns which has no significance has also we have already removed from the data and the data visualization to understand the problem. The second phase has already been completed. We then apply Artificial Neural Network as the main model to predict the classes of price and do the comparison with other machine learning algorithms like Random Forest and XGBoost

The third phase of the Project is the Data Preprocessing which requires the finding the correlation between the factors if the variables are numerical. Then we apply the machine Learning algorithms starts with the logistic regression, Random Forest and Artificial Neural Network to predict the availability class which is binary classification problem. Then we apply these same machine learning algorithms to predict the Multi classification of Predicting the class whether it is the good , affordable or expensive. Then we do the comparative analysis based on the classifiers used on both type of problem whether the problem is multi class or binary class and do the analysis and complexity analysis of the method.

## VI. PROPOSED SOLUTION

In this section, we discussed the proposed solution of our approach by classifying the classes of Price and availability by utilizing our novel deep learning model approach which is the modified version of MLP by using Python framework Keras.

### A. DATASET DETAILES

Initially, the data is too dirty as it is filled with unwanted features which is termed as noise in the field of data science. There are many independent variables which has no significance in the prediction of dependent variables. So, we have 106 features which has 20303 rows in the dataset of Airbnb in the listings of Toronto [24]. There are many columns which has no significance like scrapper id, id, picture URL, description, host picture URL, license number and other etc.

There are some categorical variables which have no significance based on the analysis of variance like country, city, access, cancellation policy, review score location etc. There are also duplicates column in the data set too neighborhood and neighborhood cleansed which are the same so take an assumption by dropping one. All these variables which are unwanted has been removed from the dataset. All the features which are were initially present in our dataset are shown in the figure.9


Figure.9: Showing the Dataset Features in the Toronto Airbnb

| S.No | Feature | Type |
|---|---|---|
| 1 | Host response Rate | Object |
| 2 | Host is super host | Object |
| 3 | Host listing count | Object |
| 4 | Host identity verified | Object |
| 5 | Neighborhood | Object |
| 6 | Latitude | Float64 |
| 7 | Longitude | Float64 |
| 8 | Is location exact | Object |
| 9 | Property type | Object |
| 10 | Room type | Object |
| 11 | Accommodate | Integer64 |
| 12 | Bedrooms | Integer64 |
| 13 | Bathrooms | Integer64 |
| 14 | Beds | Integer64 |
| 15 | Bed type | Object |
| 16 | Security deposit | Float64 |
| 17 | Cleaning fee | Float64 |
| 18 | Guests included | Integer64 |
| 19 | Minimum nights | Integer64 |
| 20 | Maximum nights | Integer64 |
| 21 | Number of reviews | Integer64 |
| 22 | Review score rating | Float64 |
| 23 | Calculated host listing count | Integer64 |
| 24 | Date difference | Float64 |
| 25 | Availability class | Object |
| 26 | Price Class | Object |

Table.1: Showing the extracted features in Airbnb Toronto Dataset

There are many missing values which are going to be handled by replacing with mean because missing values constitutes 30 percent of the data. If missing rows removed, then there might be the chance of losing an important data. Based on review dates of the first and last review, we transformed by taking the difference make the separate column named as data difference. Availability class and Price Class are the dependent variable which we are going to predict by using different machine learning algorithms. Host response rate represents the rate of the host. Host is super host represents in Boolean that whether the person is super host or not. Host listing count represents the number of hosts in the house. For example, in an apartment there are many apartments. Latitude, Longitude, is location exact, Property type, room type, accommodates, bedrooms, bathrooms, beds and bed type are related to the place and feature of the host house. Security deposit and cleaning fee are also related to the place where it is hosting. Maximum nights and minimum nights that customer can take and how many guests included in the specific price. We further removed the calculated listing host count as it is similar too the host listing count.

Moreover , the detailed statistically analysis are going to be explain in the next sub part of the same section.

### B. DETAILED DATA ANALYSIS

The most common technique to find the correlation between the Quantitative variables is Pearson Correlations. Latitude and longitude are not corelated because Pearson correlation only finds the linear correlation between Quantitative variables. Some features have strong linear relationship like accommodates and beds which is 0.84. Cleaning fee and bedrooms as we know both are corelated, so it has 0.48. zero between the variables showing no correlation while negative values representing the correlation, but one quantity decreases while other increasing. All correlations are shown in the figure.10, drawn with the help of matplotlib and Seaborn are the best python libraries to visualize the correlation.

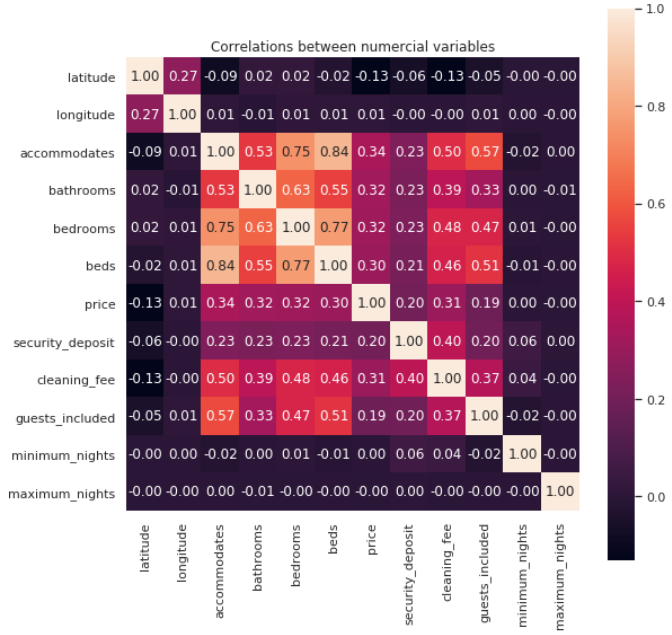


Figure.10: Showing the Pearson correlation between quantitative variables

As we know that our target variables are categorical which are Price Class and Availability class. To find the correlation between quantitative and categorical variable and categorical and categorical variable , Chi Squared technique is used which is highly based on p-value. To check the dependency which compared the actual and critical value based in the distribution of the variables and tells whether variables are independent or dependent based on the p-value of 0.05. If the value of p-value is less than 0.05 then null hypothesis must be rejected while if the value is greater than that it means, there is no dependency between variables and Null hypothesis followed with reference to the degree of freedom. To accomplish this instead of doing it manually by calculating p-values of all variables then calculate the threshold by using the degree of freedom and comparing it to see the whether the variables are dependent or not.

We write a function by setting the value of p-value of 0.95. First, we create a contingency table using crosstab and then calculate the chi squared using python library Scipy and then comparing the critical value we checked the dependency between variables by looping through all the column we get the following dependency as shown in the table. II

| Independent Variable | Price Class | Availability Class |
|---|---|---|
| Neighborhood | Dependent | Dependent |
| Host is super host | Independent | Independent |
| Host Identity verified | Independent | Independent |
| Location exact | Independent | Independent |
| Property type | Dependent | Dependent |
| Room type | Dependent | Dependent |
| Bed type | Dependent | Dependent |
| Latitude | Independent | Independent |
| Longitude | Independent | Independent |
| Accommodates | Dependent | Dependent |
| Bathrooms | Dependent | Dependent |
| Bedrooms | Dependent | Dependent |
| Beds | Dependent | Dependent |
| Security deposit | Dependent | Dependent |
| Cleaning fee | Dependent | Dependent |
| Guests included | Dependent | Dependent |
| Minimum nights | Dependent | Dependent |
| Maximum nights | Dependent | Dependent |
| Host listing count | Dependent | Dependent |

Table II: Representing a Chi-Squared Test Dependency check

### C. ALGORITHMS USED FOR PRICE CLASS PREDICTION AND AVAILABILITY PREDICTION

1. ARTIFICAL NERUAL NETWORK

Artificial Neural Networks (ANN) systems are intelligent computing systems that resemble the biological neural networks in human brains. An ANN consists the networks of connected units or nodes called artificial neurons. In an ANN, a typical artificial neuron receives the signal, process it according to activation function used to program it to send to the next artificial neurons connected to it via a connection between two consecutive nodes. One of the most popular ANN paradigms is the feed-forward neural network (FNN) and the

associated backpropagation training algorithm. Feedforward Neural Networks are the type of artificial neural networks where the connections between do not form a cycle. Feedforward neural networks were the first type of artificial neural network invented and are simpler than their counterpart, recurrent neural networks. They are called feedforward because information only travels forward in the network (no loops), first through the input nodes, then through the hidden nodes (if present), and finally through the output nodes [28].

We cannot use recurrent neural network because our data is not continuous which requires more memory but very accurate algorithm to predict the classes as well as value in regression. To fit for our model for the dataset we need to do more preprocessing of data to fit for the model. We created dummy variables for the variables Neighborhood, Property type, room type , bed type. We also created binary variable for the variable is location exact. All other continuous variables like accommodates, beds, bedrooms, bathrooms, security deposit, cleaning fee, minimum nights, guests included, and maximum nights have been scaled using scikit learn with the range from 0 to 1 based on the value. We also divided the dataset into test and train data set based on the ratio of 80 percent and 20 percent. In our modified Artificial Neural Network,

**Input Layer:** This is the initial layer of a neural network supply the input data or features to the network. We have 190 layers of input which including 175 dummy variables and 13 continuous variables. The dummy variables we created from 4 main variables which are The 15 quantitative variables are security deposit, cleaning fee, date difference , beds, bathrooms, review score rating , host response rate, minimum nights, maximum nights, guests included , number of reviews, host response rate and host listing count, accommodate and availability. We applied 500 neurons in the first layer initially.

**Output Layer:** This is the final layer which gives out the predictions. As we have 3 output layers which shows the Price is cheap, affordable and expensive by representing 0, 1 and 2

**Hidden layer:** A feedforward network applies a series of functions to the input. By having multiple hidden layers, we can compute complex functions by cascading simpler functions. We have 2 hidden layers which has 100 neurons each.

**Activation Functions:**

The activation functions used in

- Softmax
- Rectified Linear Unit (ReLU)

There are many other activation functions as an activation function but in our problem, rectifier linear units are used to save from the vanishing gradient problem if used sigmoid or tanh. Softmax used in the output layer as we have multi classification problem.

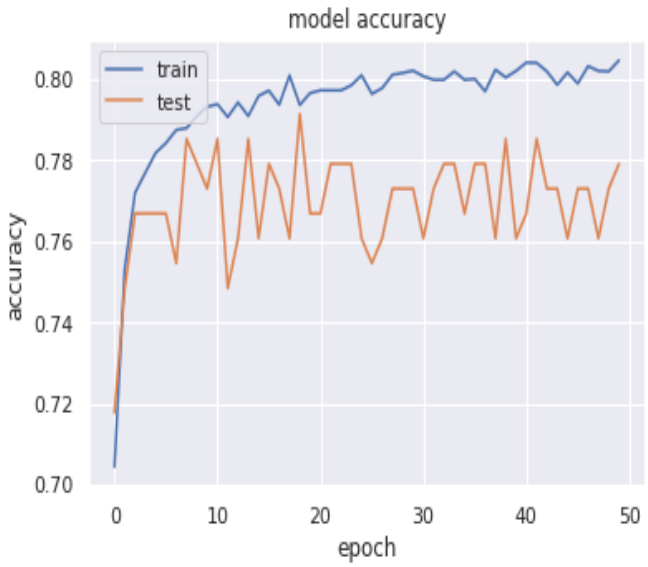


Figure.11: showing the accuracy of ANN model in train and test set

Similarly, figure.11 showing the accuracy of test set which fluctuates around 76 percent while train accuracy reaches 80 percent with 50 number of epochs. Batch size used was 32 and the validation split is 0.01.

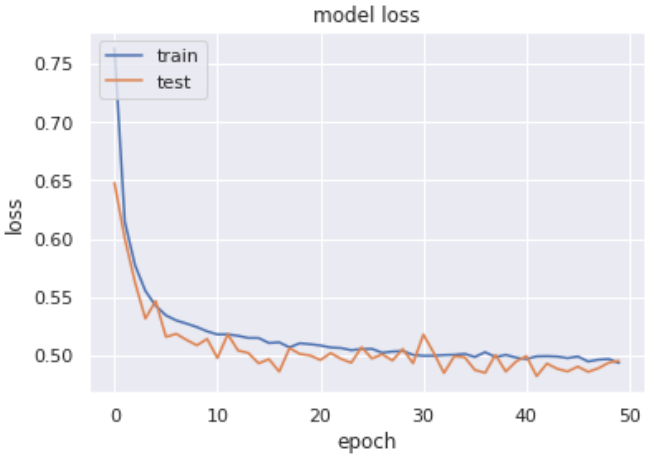


Figure.12: Showing the loss function in train and test set

The loss function in the figure which is showing in the figure.12 which is started from the values of 0.75 in the train set and 0.65 in the test set and then reduced to 0.50. As the model is overfitting because train loss function is greater than the validation and the accuracy is 78 percent initially.

**Parameter Tuning:**

To handle the problem of overfitting in our model, we use kernel initializer in the Keras's train function. We also removed 22 dummy variables which are neighbourhood_Beachborough, neighbourhood_Mansevalley,neighbourhood_MarklandWoods ,neighbourhood_Rouge,neighbourhood_TorontoIslands,property_type_Boat,property_type_Barn,property_type_Casaparticular,property_type_Cabin,property_type_Camper/RV,property_type_Cave,property_type_Cottage,property_type_Earthhouse,p

roperty_type_Hotel,property_type_Inlaw,property_type_ParkingSpace,property_type_Tent,property_type_Tinyhouse,property_type_Treehouse,property_type_Castle,neighbourhood_The Elms. After removing these variables.

We also changed the design of the ANN which is reduced to 168 input shape along with the 300 neurons applied. In each layer we applied L2 regularization of rate 0.001 which is the Ridge Regression which add to the loss function by taking the squared of the magnitude and selected the lambda of 0.001 as to remove overfitting it should not be large [29].

Similar changes have already made in the two of the hidden layers by applying 200 neurons each with drop out rate of 0.5. Also, batch size has been increased to 100 which is taking more memory and epochs increased to 300 to make our model stable. Validation split has also been increased to 0.1. for the compile part we use the same optimizer of Adam, loss function as categorical entropy because of multi classification and metrics as accuracy.


Figure.13: Showing the accuracy in the updated design

Figure.13 which is showing the model accuracy of the model with the help of updated design. The accuracy in the train set is 94 percent while the accuracy in the test set is 92 percent which is the significant improvement in the model by observing the 300 epochs.

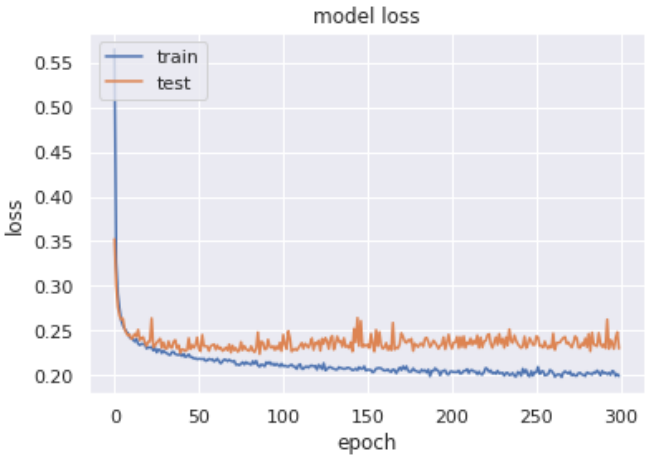


Figure.14: Showing the loss function in the train and the test model

Similarly, in the figure.14 we can observe the loss function in the test is greater than which is actual representation of the model whereas in the previous model it was reverse, so the model was overfitting. Based on the output we select and design the model. For the training loss we have 0.20 and for the test loss we have 0.24. So, no more overfitting.

We developed the artificial neural network for the Price class prediction part and utilized the same model for availability prediction as well. The final model we achieved as shown in the figure. 15 for the price prediction part where all details of the architecture have well defined and shown explicitly.

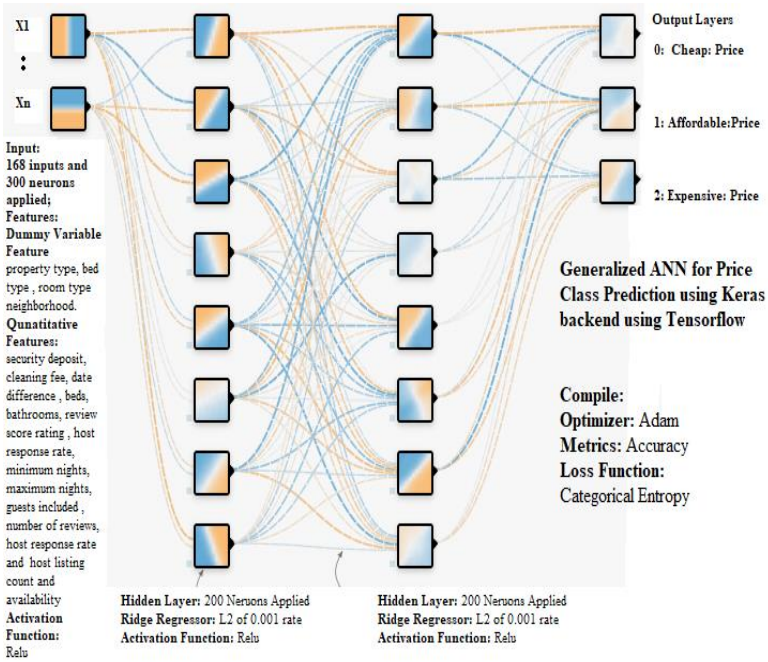


Figure.15: ANN architecture for Price class prediction

After developing the model, we also tried the same design in our other problem of predicting classes of availability high and low using the other features. With the help of TensorFlow playground [30] we draw and represent the model how it looks like physically as shown in the figure.15 and figure.16

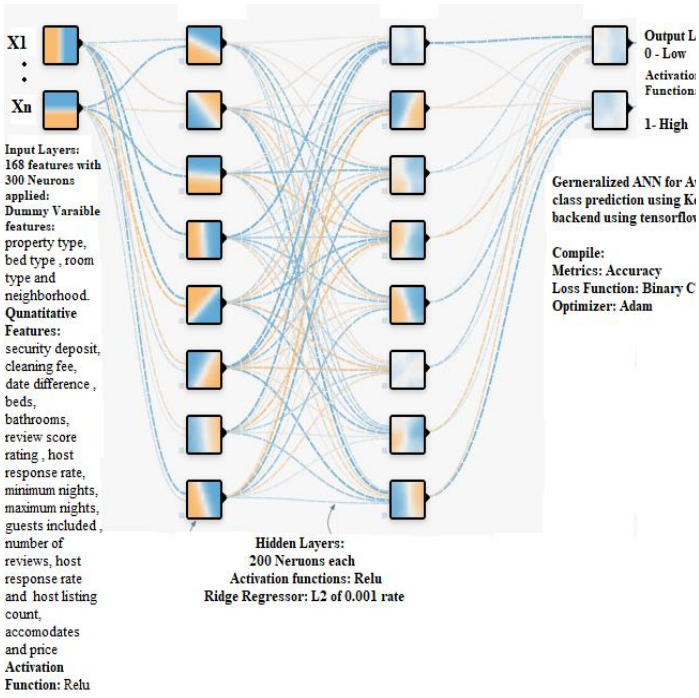


Figure.16: Showing the ANN model for Availability prediction

The only change in this design is we have price as a feature to predict availability also activation function is changed to sigmoid and loss function is changed to binary cross entropy because of binary class problem.

**Pseudo Code:**

In our model which is highly dependent on the data cleaning procedure. We removed the features which are either empty , low variance or duplicate column as these have no significance with respect to dependent variable. Then remove the symbols in the column which is creating problem for the data analysis part then we are looping through 1 to 3 in the code to run the code smoothly by initializing the counter with the help of column length. After that we created two important functions which are transforming the numerical columns to categorical column with the help of lambdas expression and apply function to apply all the rows in the column. Apply is used which is optimized way of looping through the column in pandas which is taking less time as compared to nested for loop structure. The functions are pricecol and availability.

Then we check the correlation between the numerical factors and removed the features which has 0 correlation between the factors. The correlation is between -1 and 1 so, 1 means strong correlation and increasing with increasing dependent variable while -1 showing the decreasing with increasing dependent variable.

As there are many columns which are categorical, and our dependent variables are categorical, so we apply chi square in all columns to check whether the variable are independent or not by checking the null hypothesis. The detailed steps are shown in the figure.17 as a Pseudo Code.

```
########################## PSEUDO CODE ##################################
### PROCESS OF DATACLEANING, DATA PREPROCESSING AND ALGORITHM IMPLEMENTATION ###
#########################################################################
#Initializing python notebook thorough Google Colab
Importing the data set
Checking the variance of the features in the dataset
Removing low variance, duplicates and empty features
Removing symbols '/, :,',' in the columns security deposit,
                cleaning fee and price
Removing symbols '%, :' in the column host_response_rate
Transformation of first and last review taking the difference
Replacing missing values with mean
forloop(1 through 3):
        if length of the data frame columns = 29
                deleting First Review
                deleting Last Review
        else:
        deleting Unnamed columns
        dropping missing values
        Function availability
                Pass In: passing availability column from the data frame
                if availability is greater than or equal to 180
                        return 'high'
                else return 'low'
                endif
            Call: asd(local variable) through lambda expression and apply function
                Pass Out: Availability class with high or low classes in the data frame
        Endfunction
        Function pricecol
                Pass In: passing price column from the data frame
                if price is less than or equal to 100
                        return 'cheap'
                endif
                if price is within 100 and 200
                        return 'affordable'
                endif
                if price is greater than 100
                        return 'expensive'
                endif
            Call: asd(local Variable) through lambda expression and apply function
                Pass Out: Price class with cheap, affordable and
                          expensive classes in the data frame
        Endfunction
end forloop
Checking correlation and removing the features
with no correlation between the dependent variable
for loop iteration through the columns in the data frame:
        calculating the cross table of all the columns
        applying chi squared using chi2_contingency function
        storing stat degree of freedom , expected and the probability
        setting p of 0.95 to check hypothesis
        calculating critical by giving prob and degree of freedom
        if absolute of stat is greater than critical
                Reject null hypothesis
        else:
                We cannot reject null hypothesis
        endif
        deleting local variables to release the space
end forloop
```

Figure.17: Showing the first part of the Pseudo Code and procedure of our main method

After that , all numerical features have been normalized that is between 1 and 0 so our model will not overfits. Categorical variables are transformed into dummy variables which are representing separately in each column and the features which have low variance have been removed. After that we divide the dataset into test and train based on the 20 percent to 80 percent ratio. Then we apply our model ANN and set the parameters which gives the maximum accuracy as shown in the figure.18.

```
Function str_bol
        Pass In: passing columns from the data frame
        if column is_location_exact or host_identity or host_is_superhost = 't'
                return True
        else:
                return False
        endif
        Call: a(local Variable) through lambda expression and apply function
        Pass Out: Conversion of columns is_location_exact, host_identitiy and
                host_is_superhost to boolean with True or False
Endfunction
Function pric
        Pass In: passing Price class and availability class from the data frame
        if Priceclass is cheap or Availability class is low
                return 0
        endif
        if Priceclass is affordable or Availability class is high
                return 1
        endif
        if Priceclass is expensive
                return 2
        endif
                return False
        endif
        Call: ss(local Variable) through lambda expression and apply function
        Pass Out: Conversion of Priceclass to 0,1,2 classes and
                Availability class to 0 and 1 classes
Endfunction
Forloop iteration through all columns in the data frame
        Changing type to integers
        scaling the feature from 0 to 1
end forloop
Create dummy variables for the categorical variables using concat function from pandas
forloop iteration through  categorical columns
        if value count of 1 is less than 10
        del column
end forloop
Dividing the dataset into test and train by 20 percent and 80 percent respectively
creating two sets of data one without availability and other without price
This thing helps to check model for availability and price prediction
################ ARTIFICIAL NEURAL NETWORK ######################
#################################################################
Number of Neurons First Layer: 300 * (Input = 168 Features)
Number of Neurons Hidden Layers: 200
Output Layer Neurons: 3 or 1
Number of Hidden Layers:2
Activation Function: ReLu
Optimizer: Adam
Regulizers: L2
Loss Function: Categorical Entropy or Binary Cross Entropy
Metrics:Accuracy
Kernel Initializer: Random_Normal
Dropout Rate: 0.5
Probability Selection Rate:0.5
```

Figure.18: Showing the last part of the Pseudo Code and procedure of our main method

2. RANDOM FOREST

Random forest [27] is the most powerful algorithm in classification which is the part of the terminology named as ensemble learning which is the optimization of the different machine learning algorithms to achieve good result. Random forest is the optimized version of the simple decision trees algorithm to get the best tree model to overcome the problem of overfitting. As we have already had issues of overfitting in the model when we used ANN. We select this method to see the performance of the model for comparison. So, the most important parameters in the Random forest is how much trees required to converge
the model. So, we tried with high value and then reduced to 1700 trees based on threshold to increase accuracy. We select entropy because of classification. We used the same model for price and availability prediction.

3. XGBOOST

Xgboost [28] is one of the most important and very efficient algorithms which handles high bias issue and overfitting which is also the part of the ensemble learning problem. The reason for using this algorithm in the project is best classifier for classification. For the comparison of our proposed method we also used this algorithm to compare the result.

## VII. EVALUATION

In this section, we discussed the evaluation of our findings in this paper. The previous section discussed the method details which include the implementation issues, pseudo code and architecture of our model.
The most important thing we are trying which is to improve the accuracy in the price prediction model using deep learning artificial neural network.

1. PRICE PREDICTION

The first part of the project is to predict the price classification and to do comparison with other machine learning algorithms

| TRAINING AND TEST SET DISTRIBUTION | NAME OF THE ALGORITHM | TRAINING SET ACCURACY | TEST SET ACCURACY |
|---|---|---|---|
| (Train, Test) = (16242, 4061) | Deep Learning Artificial Neural Network (Main Model) | 93 percent | 91 percent |
| (Train, Test) = (16242, 4061) | Random Forest | 81 percent | 79 percent |
| (Train, Test) = (16242, 4061) | XGBoost | 80 percent | 78 percent |

Table.III: Showing the accuracy comparison of ANN main model along with the comparison with other algorithms

Table.III showing the comparison of the other machine leaning algorithms. As the problem with our data set is high bias and low variance issue in most of the features. That's why we used ensemble learning algorithms like Random Forest and XGBoost and nonlinearity and misbalanced classes in the data set. To handle all these issues, we used Deep learning Artificial Neural which has issues and design consideration has already been discussed in the previous section. So, as we can see our main method and approach is far better than the other algorithms which are famous for classification and regression.

Similarly, the most important thing in the multi classification problem is hamming loss which gives the fractions of the classes predicted wrong to the classes all exist in the problem domain.

| NAME OF THE ALGORITHM | Hamming Loss |
|---|---|
| Deep Learning Artificial Neural Network | 0.069 |
| Random Forest | 0.20 |
| XGBoost | 0.21 |

Table.IV Showing the hamming loss comparison

With the help of Scikit learn python library we calculated the hamming loss of all these three algorithms. The nearer to zero , the more algorithm is efficient in the specific problem. So, our main method has the lowest hamming loss as compared to other.

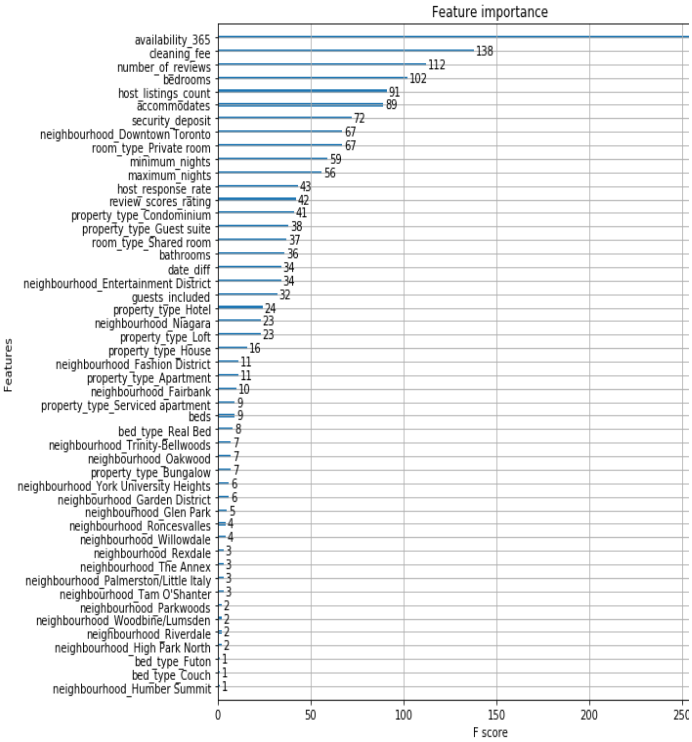


Figure.19: Showing the Feature importance in the prediction of Price Class

The top 9 features in the prediction of price class are availability, cleaning fees, number of reviews, bedrooms, host listing count, accommodates, security deposit, downtown Toronto, private room. These features are highly related to price prediction as shown in the figure. 19

## 2. AVAILABILTIY PREDICTION

The second part of the project is to predict the availability class of low and high. So, we apply the same features and same parameters are used to predict the availability of the room in the data set. The model we selected for the price with the same number of layers and same number of neurons with same regularizes of l2 are also used to predict the availability by removing availability column as discussed in the previous section.

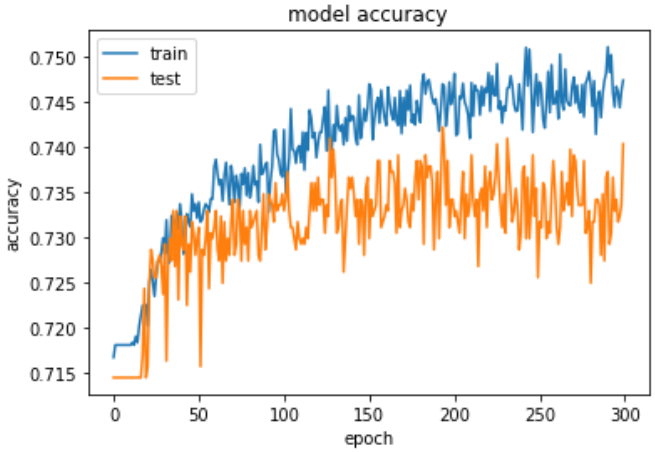


Figure.20: showing the test and train accuracy of ANN

In this figure.19 and figure.20 which are showing the accuracy and loss function respectively. The difference between train and test is low so the model is not overfitting and the loss function of test is higher than the loss function of train. So model is predicting correctly. By considering the 300 number of epochs to see the output we can understand the model loss and accuracy in the Keras.

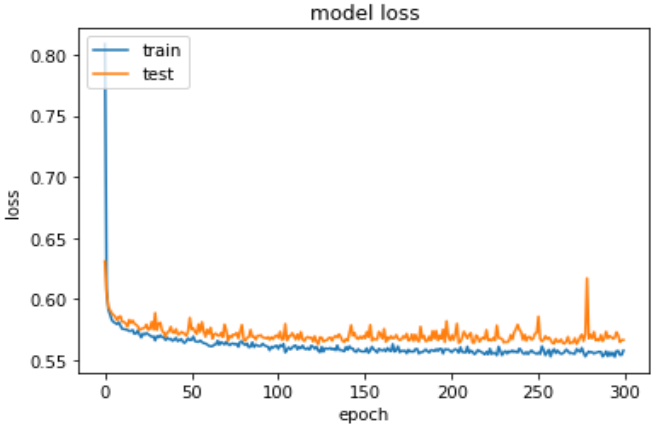


Figure.21: Showing the loss function f train and test in ANN

The most important thing in the binary classification is the confusion matrix which tells the evaluation of the classes predicted right or wrong. Table showing the confusion matrix of all the values which has been predicted high or low by using different classifiers. Now if we used the formulas of recall, precision, accuracy , specificity, accuracy and f1 score to analyze and evaluate the result more explicitly.

| CM | Artificial Neural Network | | Logistic Regression | | Random Forest | | XGBoost | |
|---|---|---|---|---|---|---|---|---|
| | **Low** | **High** | **Low** | **High** | **Low** | **High** | **Low** | **High** |
| **Low** | 2688 | 184 | 2713 | 159 | 2749 | 123 | 2625 | 247 |
| **High** | 875 | 314 | 951 | 238 | 760 | 429 | 681 | 508 |

Table.V: Showing the confusion matrix of all algorithms

If we compare the algorithm of our main approach of deep learning with other algorithms, So ANN has good accuracy of 74 percent but has poor performance as compared to Random Forest and XGBoost algorithms as they have high F1 score, high accuracy , high precision and high recall.
Another important thing in the binary classification is that we can draw the ROC curve to compare the different classifiers to check the performance of different algorithms

| Algorithm | Accuracy | Precision | Recall | Specificity | F1 Score |
|---|---|---|---|---|---|
| ANN | 0.74 | 0.63 | 0.3 | 0.94 | 0.4 |
| Logistic Regression | 0.726 | 0.6 | 0.2 | 0.94 | 0.3 |
| Random Forest | 0.782 | 0.77 | 0.36 | 0.95 | 0.5 |
| XGBoost | 0.77 | 0.67 | 0.427 | 0.91 | 0.52 |

Table.VI: Showing the evaluation parameters from confusion matrix

By considering the ROC curve for evaluation of classifiers for the specific binary problem which is based on false positive and true positive rate. As shown in the figure.22 ROC curve for all four classifiers we used in the prediction of availability. So XGBoost and Random forest are the best methods for the prediction of Availability

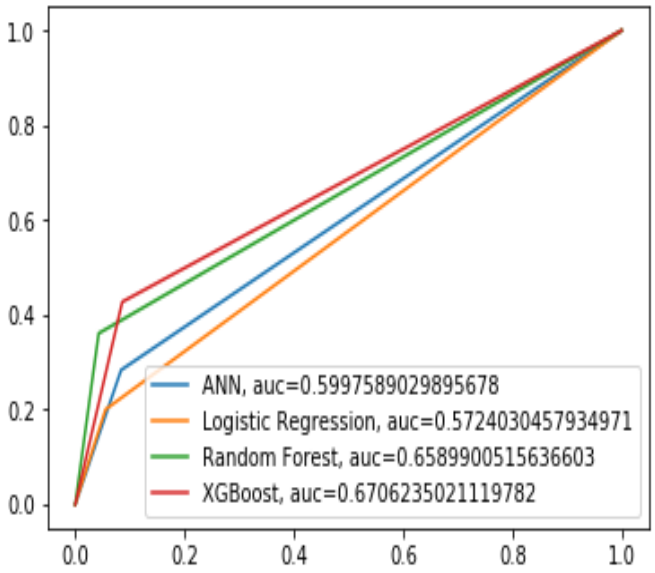


Figure.22: Showing the ROC curve for different classifiers

Now , the most important thing is what could be features important to predict the availability class. So, the features which are the most important are shown in the figure.23


Figure.23: Showing the feature importance in the prediction of Availability class

We can see that the price, date difference , Number of reviews, cleaning fees, host listing count, review rating score, maximum nights and minimum nights, security deposit and host response rate are the top features which are important in the prediction of Availability class.

## 3. COMPLEXITY ANALYSIS

The software to run Python Jupyter Notebook scripts is google colab which has 12 GB of RAM , fast processor and GPU as we are running Neural Networks which is taking a lot more time during train so we need a GPU and also using google colab give built in libraries installation, you do not need to install python libraries as most of them have already been installed, so only need to import the libraries. Because installing libraries specially TensorFlow , keras and SciPy are little bit complicated in python when you are installing in your local machine.

| EPOCHS | BATCH SIZE | NUMBER OF ITERATIONS | TRAINING ROWS |
|---|---|---|---|
| 1 | 100 | 147 | 14617 |
| 100 | same | 14700 | 14617 |
| **Total Parameters Used for training:** 40,062 | | | |

With the help of good manual of scikit learn where the generalized complexity is given by the formula

$$\mathbf{O(n, m, h^k, o, i)} \quad \text{----------(1)[26]}$$

Where **n** is the number of inputs applied to the first layer, **m** is the number of training samples, **h** is the number of layers, **o** is the output neurons and **i** is the input neurons.
So, in our model which has the architecture in price and availability prediction we have the following values.

$$\mathbf{O(168, 14617, 2^200, 3, 300)} \quad \text{----------(2)}$$

So, the equation (2) is showing the whole iterations of our model.

## 4. EFFICIENCY ANALYSIS

Our model is taking 600 seconds with the google Colab which has built in GPU, but it is taking less time when you run with good GPU and 32 GB of RAM as 12 GB RAM provided by google Colab is not enough.

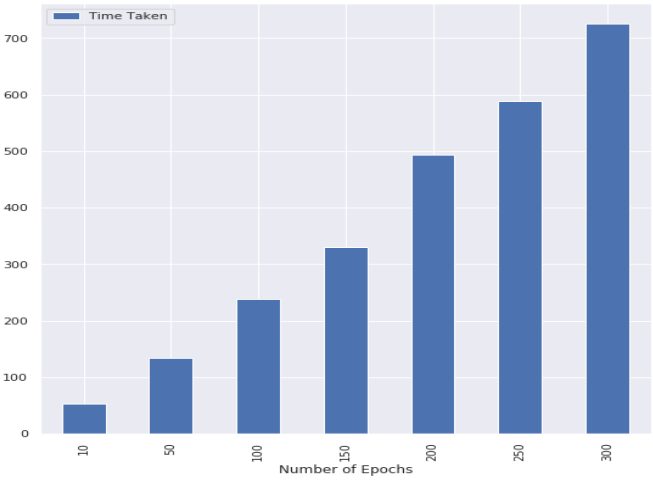


Figure.24: Showing the time taken by running model with number of Epochs

By the bar plot shown in the figure.24 as we are increasing the number of Epochs which means increasing iteration which in turn increasing time taken to run the model. The model we proposed is taking around 700 seconds to run and give prediction. The lowest time taken with epoch are 10 which is taking around 50 seconds. So, Time is linearly increasing as we are increasing the number of epochs.

## VIII. DISCUSSION

Now the question arises what are the most important factors which are highly dependent on the price and availability. Also, if we discuss the significance of the model. The most important thing whether model is working properly or not because in deep learning neural network there are many arises foe example overfitting when the difference between train and test is too much or loss function of the train is much lesser than train or significantly greater than test loss function. Similarly , if increase the number of hidden layers, the problems might occur are exploding gradient or vanishing gradients which has already taken care of by initializing kernel normal in the Keras model to handle these issues. Then our model overfits which is the most common method in deep learning models. To handle this, we initialize the L2 regularization in the model. So, our final model did not overfits in neither Price prediction nor in availability prediction. There are also other methods to reduce overfitting which include reducing the number of layers, number of neurons, applying L1 and L2 regularization, applying different drop out weights on different layers. Good fit of the model in machine learning terms if the problem is classification is that when the difference of test and training accuracy is very low but should not be equal because if they are equal then model is underfitting. So, in both application of the problem whether in price and availability our model accuracies between test and train are around 2 percent which looks good fit model. Also, if we discuss what would be the implications of the model and benefit to the world, that is another question. So, this feature if we add in the Airbnb website, this feature is not only helpful for the customers as Price classes are created based on the customer average salaries in Ontario but also, this feature is giving profit to the company , Airbnb by increasing the number of customers and customers good feedback will significantly improve the business and give a significant revenue for the company. This model will behave like a win-win situation for both customers and the firm.
To wind up the discussion, it is proved that using Artificial Neural Network using in the correct way will result in improving the model in any type of data set as awe achieved wit h 91 percent accuracy, but which requires lot of tuning the parameters by applying different combinations to achieve good accuracy. The main reason for using ANN is because in most of the world real data is nonlinearity which makes the people to use ANN in using real world datasets.

## IX. COMPARISION WITH RELATED WORK

In short, if we compared the other work most of the people work used spatial analysis of Airbnb in city or country level as discussed in the comparative analysis and related work section. In related work which motivated us to improve model is found

on Kaggle Airbnb Price Prediction [32] in Seattle but people achieved very poor Mean Absolute Error [31] which is around 0.27 . So , instead of doing regression analysis to improve the model in Price prediction, we change the domain to classification along with the new feature of predicting availability. By doing this we achieved 92 percent accuracy in the test set which shows our work novel.

## X. CONCLUSION

To put in the nutshell, the accuracies of price prediction is very low on the web used by different people on the Airbnb different cities dataset. Instead of Predicting the exact value of price to see the model accuracy, we transformed the prices into classes based on hypothesis and then predict the classes using our deep learning model not only to predict the class but also to predict the availability because these two are the most important factors and achieved 92 percent accuracy in the Price prediction which is greater than other algorithms like random forest and Xgboost. The hamming loss of our model is also lesser than other methods and achieved 0.069 as the nearer to zero , the more accurate the model result is. Same model is also applied to predict the availability and compared with other algorithms but for the availability , less accuracy achieved due to the more tuning of parameters required.

The most important factors for the prediction of price class are availability , cleaning fee, number of reviews, and number of bedrooms. So if a person is interested to rent aa room in Toronto, the most important thing is to see the availability whether it is available or not then check the cleaning fee which will also help to understand the price and similarly, number of reviews so the more reviews the more price and also increasing number of bedrooms also increase the price. By introducing new feature in the Airbnb system will also increase the profit for the company as it fulfils the needs of the customers as the people who are renting the house, they have specific range in their mind. Increasing the customers booming the business.

Similarly , for the availability the most important factors are price, number of reviews, cleaning fee, and the period of the time of the reviews which could impact on availability of place in Toronto.

## XI. FUTURE WORK

There are many directions to improve this novel work. First is to improve the model by selecting the features as we used Pearson correlation method along with analysis of variance which is used to filter the features but non linear dimension reduction methods like Multi-dimensional scaling, isometric feature mapping locally, linear embedding, Hessian Eigen mapping, t-Distributed stochastic Neighbor embedding and Auto encoders could be used.

For the algorithm part as we used deep learning neural network, we cannot use convolutional neural network as there are no images in the dataset also use of recurrent is also not possible as our data is not time series so every row is dependent on the previous one. To improve the model , there are methods like Grid search and Auto Keras which are the latest methods for the Hyper Parameter tuning to improve the model of Price and Availability prediction.

As the time is constraint but the model can further be increased to the National level by combining the data of all the cities in Canada as we only focused more on Toronto.

## XII. ACKNOWLEDGEMENT

I would like to thank my supervisors **Prof. Ioannis Lambadaris (Department of Systems & Computer Engineering)** and **Prof. Omair Shafiq (School of Information & Technology)** for their valuable guidance and motivation which was essential for the progress and completion of this research work and its implementation. Throughout the whole time from beginning to end their valuable guidance, tips and advises helped me in completing this task, and provide a clear and detailed picture of the topic which is very helpful and guide me to accomplish this novel work.

## REFERENCES

[1] Daniel Guttentag, Stephen Smith, Luke Potwarka, Mark Havitz,"Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study", Journal of Travel Research, Vol 57, Issue 3, pp. 342-359, 2017.

[2] The Statistics Portal. Number of Airbnb users in the United States from 2016 to 2022 (in millions). retrieved from https://www.statista.com/statistics/346589/number-of-us – airbnb-users/. 2018.

[3] Paridhi Choudhary, Aniket Jain, Rahul Baijal, Unravelling "Airbnb Predicting Price for New Listing", Cornell University Library, 2018.

[4]. Ki-Hong Choi1, Joohyun Jung2, Suyeol Ryu3, Su-Do Kim4 and Seong-Min Yoon1," The Relatrionship between Airbnb and the Hotel Revenue: In the Case of Korea", Indian Journal of Science and Technology, Vol. 8, ISSN : 0974-6846, 2015

[5]. Laurent Son Nguyen, Salvador Ruiz-Correa, Marianne Schmid Mast, Daniel Gatica-Perez," Check Out This Place: Inferring Ambiance From Airbnb Photos", IEEE Transactions on Multimedia, 2018

[6]. Longhua Guo, Jianhua Li, Jie Wu, Wei Chang, Jun Wu, "A Novel Airbnb Matching Scheme in Shared Economy Using Confidence and Prediction Uncertainty Analysis", IEEE Access, 2018

[7]. Qian Zhou, Yang Chen, Chuanhao Ma, Fei Li, Yu Xiao, Xin Wang, Xiaoming Fu, " Measurement and Analysis of the Reviews in Airbnb ", 2018 IFIP Networking Conference (IFIP Networking) and Workshops, 2018

[8]. M. Abdar, K. Lai, Ny Yen, "Crowd Preference Mining and Analysis Based on Regional Characteristics on Airbnb", International Conference on Cybernetics, IEEE, 2017

[9]. Czesław Adamiak , Barbara Szyda, Anna Dubownik, David García-Álvarez ,“Airbnb Offer in Spain—Spatial Analysis of the Pattern and Determinants of Its Distribution”,” IEEE Transactions on Big Data”, 2019

[10]. Giovanni Quattrone, Andrew Greatorex, Daniele Quercia, Licia Capra and Mirco Musolesi, “Analyzing and predicting the spatial penetration of Airbnb in U.S. cities”, EPJ Data Science, 2017

[11]. Hanna Leea, Sung-Byung Yangb, Chulmo Koob,” Exploring the effect of Airbnb hosts' attachment and psychological ownership in the sharing economy”, Tourism Management, 2018

[12]. Wang Min, Li Lu, “Who Wants to live like a local? An analysis of Determinants of Consumers’ intention to choose Airbnb”, ICMSE, IEEE, 2017

[13]. Raza Hasan, Mehdi Birgach, “Critical success Factors behind the Sustainability of the sharing Economy”, International conference on Software Engineering Research management and application, 2016

[14] Kun Wang , Chenhan Xu , Yan Zhang, Song Guo ,Albert Y. Zomaya, “Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid”, IEEE Transactions on Big Data, Vol. 5, 2019

[15] Subhajit Sidhanta, Wojciech Golab, Supratik Mukhopadhyay, “Deadline-Aware Cost Optimization for Spark”, IEEE Transactions on Big Data, Vol. 6, 2018

[16] Malav Shastri, Sudipta Roy and Mamta Mittal, “Stock Price Prediction using Artificial Neural Model: An Application of Big Data”, EAI Endorsed Transactions on Scalable Information Systems, 2019

[17] Ying Yu, Yirui Wang, Shangce Gao,Zheng Tang, “Statistical Modeling and Prediction for Tourism Economy Using Dendritic Neural Network”, Computational Intelligence and Neuroscience, 2017

[18] Mathias Longo, Matías Hirsch, Cristian Mateos, Alejandro Zunino, “Towards Integrating Mobile Devices into Dew Computing: A Model for Hour-Wise Prediction of Energy Availability”, MDPI, 2019

[19] Sheikh Mohammad Idrees , M. Afshar Alam, Parul Agarwal, “A Prediction Approach for Stock Market Volatility Based on Time Series Data”, IEEE Access, 2019

[20] Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, Sheraz Ahmed, “DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series”, IEEE Access, 2018

[21] Tongton Yuan, Weihong Deng, Jiani Hu, Zhanfu An, Yinan Tang, ”Unsupervised adaptive hashing based on feature Clustering”, Elsevier Neurocomputing, 2019

[22] Neha Bharill, Aruna Tiwari, Aayushi Malviya, “Fuzzy Based Scalable Clustering Algorithms for Handling Big Data Using Apache Spark”, IEEE Transactions on Big Data, Vol. 2, 2016

[23] Ellie Ordway-West, Pallabi Parveen, Austin Henslee, “Autoencoder Evaluation and Hyper-parameter Tuning in an Unsupervised Setting”, IEEE International Congress on Big Data, 2018

[24] Dataset retrieved from: https://insideairbnb.com/get-the-data.html

[25] Sharma, S. (2017). activation Functions: Neural Networks. Retrieved from https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[26]L2 regularization concepts retrieved from www.towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

[27] Understanding random forest retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[28] Xgboost introduction data retrieved from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[29] Neural Network complexity data retrieved from https://scikitlearn.org/stable/modules/neural_networks_supervised.html#complexity

[30] Tensorflow playground online retrieved from https://playground.tensorflow.org

[31] Difference between RMSE and MAE retrieved from https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

[32] Airbnb Price Prediction in Kaggle retrieved from https://www.kaggle.com/naamaavi/airbnb-price-prediction-regression-project