Analysis on New York Airbnb Rent

CS513 final project

Team Members:

Name: Qianyi Zhang, CWID: 10455276

Name: Zichong Wang, CWID:10464881

Name: Zheng Liu, CWID:10455895

Name: Ziming Zhang, CWID:10455301

Abstract

Section I introduces some basic background knowledge of the project, and then Section II outlines the problems of the project and what to accomplish. Section III discusses the technology used in the project and the results obtained. Finally, Section IV summarizes the project.

I. Introduction

Customers need to consider many factors when looking for a suitable house, such as room type, house price, location and so on. Therefore, finding a house is very difficult. Customers have to compare many options. There is also an inseparable connection between these factors, from which we can analyze these factors to classify and predict the results we want. For example, we can find the characteristics of house prices from other factors.

II. Problem Description

We intend to use Machine learning models to help customers get the price of a house in the desired location and room type. And show how different variables affect housing prices. At the same time, it can also show homeowners what are the factors that affect house prices. Here we have obtained a set of Airbnb rent data in New York. We intend to analyze and predict the price of the house from the location of the

house, such as latitude, longitude and neighbor groups. Of course, we also need to consider the room type. We perform data preparation, data analysis and data modeling based on the original data obtained. And we use different data mining methods and prediction models to predict housing prices.

III. Techniques Used

Data Modelling:

We have our final dataset after cleaning. We now must start modelling- Predicting the price of the Airbnb in NYC. Most of the time in Regression and Classification problems, you run your model with the available values and check the metrics like accuracy of the model by comparing observed values with true values. We train the model with the training set in a way that it can be applied to testing set.

KNN

Helps in classifying the data based on how its nearest neighbor is classified(according to Euclidean distance), when measuring distance, one or more attributes can have very large values, relative to the other attributes. In order to balance the weight of each attributes, we use Min-Max Normalization to normalize our data.

$$\text{Min - Max Normalization} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Related Features : Latitude Longitude room type reviews availability.

According to the accuracy, K = 50 can get the best accuracy.

KKNN

Performs k-nearest neighbor classification of a test set using a training set. For each row of the test set, the k nearest training set vectors (according to Minkowski distance) are found, and the classification is done via the maximum of summed kernel (k value) densities. In addition even ordinal and continuous variables can be predicted. From the predicted values, find the error rate and accuracy.

Related Features : neighbourhood_group neighbourhood room type.

K = 3 => 60.6% accuracy

K = 10 => 63.6% accuracy

Naive Bayes

Naive Bayes is a family of algorithms that all share a common principle: every feature is independent of each other. A Naive Bayes classifier considers each of these features to contribute independently to the prediction, regardless of any correlations

between features. This is, of course, almost never true, which means that the applications of this model are very situational. However, it often gives good results, showing better results than other models that do not have their hyper-parameters fine-tuned. It's a pretty simple model.

Related Features : neighbourhood_group, neighbourhood, latitude, longitude, room type.

Accuracy : 64.53%

Decision Tree - CART

Classification And Regression Tree is a simple technique to fit a relationship between numerical variables partitioning the target variable by a range of values of the explanatory variables. This function fits and graphs a cart model with a previous separation of training and testing datasets.

Related Features : latitude, longitude, room type.

Accuracy : 65.75%

SVM

The next model for fit is a support vector machine(SVM) model. Basically, an SVM constructs a hyperplane or a set of hyperplanes that have the largest distance to the nearest training data points of other classes. Choose a radial kernel with proper gamma and cost values here to optimize the performance of SVM. The SVM model object is a list presenting basic information about the parameters, number of support vectors, etc. The number of support vectors depends on how much slack we allow when training the model. If we allow a large amount of flexibility, we will have many support vectors.

Related Features : neighbourhood_group, neighbourhood, room_type.

Accuracy : 65.9%

Artificial Neural Network

Neural Networks are a machine learning framework that attempts to mimic the learning pattern of natural biological neural networks. Biological neural networks have interconnected neurons with dendrites that receive inputs, then based on these inputs they produce an output signal through an axon to another neuron. We will try to mimic this process using Artificial Neural Networks (ANN). Fits a single hidden layer ANN model to input data x and output data y. We choose Min-Max Normalization, and 15 hidden node for our model.

Related Features : neighbourhood_group, neighbourhood, room_type.

Accuracy : 68%

Random Forest

When working on a classification problem, it's almost always a great idea to start with RandomForestClassifier; not only it's a pretty good classifier for many problems, but it also allows you evaluate what was the impact of each feature on the prediction. This is precious information, as one can build better models by removing undesirable

features. There are other ways of figuring out which features are important to our model as well.

Related Features : latitude, longitude, room type.

Accuracy : 67.83%

C50

Simple to understand, interpret, visualize. Decision trees implicitly perform variable screening or feature selection. Can handle both numerical and categorical data. Can also handle multi-output problems. Decision trees require relatively little effort from users for data preparation. Nonlinear relationships between parameters do not affect tree performance.

Related Features : latitude, longitude, room type.

Accuracy : 68.1%

IV. Conclusion & Future Work

Conclusion

Machine learning models are as good as the data you feed, and more data would strengthen the model. While some level of attrition is inevitable it should be kept at the minimum possible level using above solution. Besides the amount of the data, the data preprocessing is very important, we need to filter extreme data, wrong data and invalid data, which has a big impact on our model.

Based the most important features to the least important features we can identify what are the main causes for attrition.It helps to understand the key variable that influence the house rent.

We have considered every selective group of features to identify what works best for our dataset and analyzed our model solution.

Future Work

More machine learning techniques can be applied to our model, more data processing we can do to get a better data and to get a better performance. Advanced machine learning models and selective features can help improve our predictive analysis. Using ensemble model to breakdown complex structure into critical features that most related to house rent.

V. Related links

GitHub Repository: https://github.com/yeezycheung1997/-CS513