DSC 412 Fall 2024 Exploring Machine Learning- Project Proposal

Student Name: August Pallesen
Student ID: 200618081
Email: apalles@ncsu.edu

*** This project will be combined with the project I am doing in *CSC 422 Automated Learning and Data Analysis.* The project requirements for this project can be found on the next page ***

Proposal:

As an foreign exchange student, I would like to travel as much as possible while I am here. Doing so is expensive and time consuming, f.ex. finding the right airbnb to stay at, while still traveling on a budget is difficult yet important. So why not create a machine learning model to predict the price of a given airbnb. This is the aim of this project, namely to build a ML model that within a relatively high degree of accuracy can predict the price of a given air bnb listing in New york (where I will be going soon, and spent a lot of time trying to find the right place). This will hopefully help other like minded people who want to travel to new york. In the development of this project, several challenges may arise, such as flawed or incomplete data, null values, large quantities of data. The novelty of this project lies in the problem, as I wont be using any other machine learning techniques or approaches than what is already available.

Plan:

The project will be developed using the dataset: New York Airbnb Data Mining from kaggle. The dataset consists of approx. 25k entries, each detailed through 22 features. The goal is to analyze the data through careful data preprocessing, and then to apply a regression model, in order to find any trends or patterns in the data that could help us gain insight into the volatile airbnb listing market in New york. The accuracy of the model will be measured through performance metrics such as confusion matrices, RMSE, and/or other accuracy metrics. Since the dataset is from kaggle, other data scientists may have tried to develop a similar model, which could prove useful as a baseline for improvement down the line.

CSC422 Fall 2024 - Project Proposal

Student Name: Isaac Adams, August Pallesen
Student ID: 200524264, 200618081
Email: jiadams2@ncsu.edu, apalles@ncsu.edu

1 - Chosen Project

Title: New York Airbnb Data Mining
Category: Exploratory Data Analysis
Dataset: Kaggle Dataset
Difficulty: ***

2 - Project Abstract

With this project we have divided our core objective into 2 main deliverables our final model will accomplish, price prediction and demand forecasting. Our price prediction algorithm aims to predict the price of an Airbnb listing based on its features (e.g., location, room type, number of amenities) and the demand forecasting algorithm aims to forecast demand for different neighborhoods or types of Airbnb listings.

3 - Proposed Methodology

In order to develop a model that truthfully captures the trends of airbnb's in New York, it will be necessary to ensure that bias is not introduced in the model. This should be done through careful and rigorous data collection, preprocessing and training. Exploratory Data Analysis (EDA) will be performed on data to gain statistical insight and to better understand the underlying data. The choice of algorithms will depend on the problem type; for example, we will probably use a regression model for predicting prices, meanwhile classification models are more appropriate for predicting listing types. Cross-validation techniques, such as k-fold cross-validation, will be used to assess model performance and avoid overfitting. We will use hyperparameter tuning to optimize algorithm parameters, in order to improve model accuracy and efficiency. Finally we will measure the performance of the model using some evaluation metrics like RMSE and classification accuracy or confusion matrix.

4 - Expected Outcomes

In this project we anticipate being able to yield insight findings from the data, such as being able to predict prices and demand forecasting. As mentioned above we will try to measure the quality of our results through performance metrics such as RMSE and classification accuracy or confusion matrix and maybe more.

5 - Timeline

- Week 1: Data Collection and Preliminary Exploration
- Week 2: Data Collection and Preliminary Exploration
- Week 4: Data Cleaning and Preprocessing
- Week 5: Data Cleaning and Preprocessing
- Week 6: Initial Analysis or Model Training
- Week 7: Further Analysis or Model Evaluation
- Week 8: Further Analysis or Model Evaluation
- Week 9: Model Evaluation and Performance Testing
- Week 10: Summarize Findings and Prepare for Presentation

6 - Possible Challenges

In analyzing the New York Airbnb dataset, several challenges may arise. Data quality issues, such as missing or inconsistent entries, need to be handled by means of thorough cleaning and preprocessing. Handling large volumes of data may require a more efficient data processing approach. Integrating diverse data sources and accounting for the dynamic nature of the market through time-series analysis could also be a problem. Addressing these challenges will help us ensure a more accurate model that hopefully yields insights into the New York Airbnb market.