**IST 5535: Machine Learning Algorithms and Applications**

**Deliverable 1: Project Proposal**

**Group 2: Tanner Fry, Sayantan Majumdar, Daren Liu**

**Problem description and Research Questions**

This project aims to build a machine learning model to predict the rental price and which factors might have the most impact on rental prices in New York. It also deals with comparing different machine learning models to assess their performance on the dataset. In this regard, the following research questions need to be answered:

1. Which predictors are statistically significant for predicting the house rentals?
2. Which machine learning model works best for this data set?

**Dataset (source, number of rows, number of columns, response variable, etc)**

This dataset can be found at <https://www.kaggle.com/sab30226/zillow-rents-2020>. It contains 22 explanatory variables (columns) that detail almost every aspect of Rent prices for apartments in New York listed on the site Zillow.com. There are 6955 observations (rows) in this dataset related to housing rental data in New York including yearBuilt, city, postal code, address, price, area, and more.

**Regression and Classification Analysis**

The New York rental price dataset is a regression-based dataset where we need to predict the rental price based on the explanatory variables In this project, we are dealing with a multivariate regression problem that is concerned with both prediction and inference.

**Potential problems and Challenges**

The major challenge in this project is the data cleaning part. This dataset contains some redundant columns like “real estate provider”, ”payment period”, “is listed” and other columns wherein there are missing data as well. This noise may affect excluding the unrelated variables. Moreover, the “facts and features” column is an aggregated one or non-atomic and suffers from redundancy due to the columns “bedrooms”, “bathrooms”, and “area” which already exist. Similarly, “price” is also a non-atomic column where the units are specified directly along with the values. In addition, inappropriate categorization of the rental price may lead to a high variance of the regression result.

**DATA APPENDIX**

|  |  |
| --- | --- |
| yearBuilt | The year of the rental properties built |
| url | Zillow URL for certain properties |
| longitude | Longitude of the rental properties in geographic degrees |
| homeStatus | Defines if the property is available for rent, sold, pending, etc. |
| real estate provider | Provider of the rental properties |
| paymentPeriod | Rent payment period |
| isListed | Whether the property is listed on Zillow |
| relaxed | Whether the rental place has been relaxed |
| postal\_code | Postal code of the rental place |
| title | Title on the Zillow website for certain properties |
| address | Rental address |
| facts and features | Number of bedrooms and area |
| currency | Type of payment currency used |
| commute | The distance of the property to the nearest public transport system |
| price | Monthly rental price in USD |
| latitude | Latitude of the rental properties in geographic degrees |
| city | City of the rental properties |
| homeType | Classifies the rental properties into multi-family, apartment, single-family, townhouse and condo. |
| state | State of the rental properties |
| bedrooms | Number of the bedrooms |
| bathrooms | Number of the bathrooms |
| area | Area of the rental properties in sq. ft |