Analysis of the Housing Market in Two US Cities from 2020-2022

Real Estate Investment Group

**Data Science Capstone Project   
Launch Report**

Date:

01/17/2023

Team Members:

Name: Nikhil Muthuvenkatesh

Name: Seyi Oyesiku

Name: Luqing Qi

Name: Rohan Ukkalam

**The System/Product**

**System/Product Name:** East Coast Real Estate Price Predictor

**Introduction:**

Since the COVID-19 pandemic in 2020, real estate pricing has been quite volatile. The initial lockdowns led to a large drop in property values across the country, but shortly thereafter there was a huge spike that has continued up until now. This project will implement a model that is trained with real estate data over the last two years in order to predict the future prices for housing. The goal of this capstone project is to correctly predict when someone that is interested in purchasing real estate should make a purchase of a home in a city on the East Coast, such as Philadelphia and New York.

The motivation behind this project comes from the team members. We are all graduate students who are interested in purchasing real estate in the future, and we would like to know what to expect from this market in the future. It could also potentially be useful for banks that give out home loans, mortgage companies, and for escrow companies. In addition, anyone that is interested in purchasing a home in one of the cities that we use for analysis in the next couple years would find this information useful.

The biggest deliverable of this project is the best machine learning model to use to accurately predict future house pricing in Philadelphia and New York. The goal of this project is to find the best machine learning model that is able to predict house pricing.

**Highlighted Features:**

* Model that is able to accurately predict house pricing in these two cities within the next couple months
* We will be implementing a logistic regression, decision tree, and a neural network to find out which of these 3 models would work the best for us.
* The training set will include real estate data from both cities from the time period of 2020-2022
* For Loan-Level Public Use Databases(PUDBs), we are trying to find the most critical variables influencing the interest rate and borrowing amount. The potential methods we are considering now include Regression Model, Tree-Based Model, Neural Network, and KNN. We also tried to apply an unsupervised learning method to find the potential relationship with the mortgage borrower—for example, Race group, Income level, and credit score.
* For Rolling Sales Data in NYC, we are trying to find the most important variables that influence the sales price in New York City. The potential methods we are considering now include Regression Model, Tree-Based Model, Neural Network, and KNN. We are also trying to use the K-Means method to find potential relationships in year built and square footage.
* End product will include visualizations for laymen to better understand the trends for the past few years
* End product will include several areas/location of interest

**Sponsor or Proxy User:**

Potential users include mortgage companies, banks, investors, and people searching for homes.

**Issues:**

We have resolved all issues with the data acquisition for this project. Our biggest challenge right now is selecting the correct prediction model that will give us the best results. We will be trying a few different supervised learning methods such as logistic regression and a decision tree because we have a target variable: price. We are also looking into several new models that we do not have much experience in due to their promising natures, but will have to be careful in approaching them.

**The Team**

**Team Name:** Real Estate Investment Group.

**Team Members and their specialties:**

Nikhil Muthuvenkatesh- Intermediate level of programming in Python, background in sales/engineering to help market the final product, some knowledge of the real estate market

* Would like to work on the model selection research, and implementation of various models chosen for our analysis
* Would like to work on coding the training and test set split

Rohan Ukkalam - Intermediate level of programming in Python, OOP, and Machine Learning.

* Would like to work on the model selection research, and the implementation and coding of various models - specifically the Neural Network - chosen for the analysis
* Would like to work on report writing

Seyi Oyesiku - Intermediate level programming in python, with background in Telecommunications and supply chain management.

* Would like to work on model selection and the implementation of the various models that will be chosen for our analysis.
* Would like to work on reporting as well.

Luqing Qi - Intermediate level programming in python, R , and SQL, background in Health Care industry.

* Would like to perform Predictive Analysis on Loan-level Public-Use Database
* Would like to perform Predictive Analysis on Rolling Sales Data at NYC.

**Team Communication:**

The team will continue with our communications over iMessage and Zoom. Because everyone on the team lives in very different locations from campus, we have found it best to meet over Zoom or Microsoft Teams. Since everyone is in the in-person class this term, we may have a few in person meetings throughout the term.

**Team Issues:**

One member of the team is currently leading a busy schedule, which might lead to scheduling conflicts - this will be remedied via stringent and punctual team meetings, such that this team member will be able to schedule comfortably around said meetings.

**Plan**

Our initial plan is to tidy up our current dataset, remove outliers and NaN values, before partitioning it into training and testing sets as required for the models we plan to use. While we have several models in mind, we’re currently researching other potential models that might be better suited to the nature of our project. After finalizing several models, we will train them, test them, evaluate whether the results they produced are reasonable, and then test the models against each other to find the best model.

Table of Contributions

The table below identifies contributors to various sections of this document.

|  | **Section** | **Writing** | **Editing** |
| --- | --- | --- | --- |
| **1** | **Project** | Nikhil Muthuvenkatesh | Nikhil, Muthuvenkatesh, Rohan Ukkalam |
| **2** | **Team** | Seyi Oyesiku, Luqing Qi | Rohan Ukkalam |
| **3** | **Plan** | Rohan Ukkalam | Seyi Oyesiku |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.