**Milestone 3 – Preliminary Analysis**

**Housing Market Prediction Analysis**

DSC630 Predictive Analysis

Sri R Sankaranarayanan

**Introduction :-**

Purchasing a house is a big decision in a person’s life and needs a considerable amount of thought and research. As a first-time amateur home buyer, I am looking to develop this project on my own personal interest to see how the housing prices in the near future is heading in various given metro cities in the USA. As we all know the current pandemic has driven the housing market so crazy and it is simply unaffordable for many young American families to dream about owning their first time, especially in cities like Dallas, Austin, Denver, Houston, Phoenix and few other major metro cities.

**Business Problem :-**

House price prediction is a significant financial decision for individuals. One would want to buy their first home or invest on real estate properties at the best rate and minimum risk and would like that to be the best investment for future. We are aiming to predict the housing price based on the historical Sale Price, Median Income, Number of Schools and their Rating, Crime Rate, Unemployment Rate, Number of Hospitals and their ratings.

**Scope/problem Statements:-**

1. What does the analysis/model building tell you?
2. What are the recommendations?
3. What will be the predicted price per square foot of a home in a given zip code/ county in the next few months?
4. Predict the price based on crime rates, schools and other information/metrics provided by Zillow for a zip code/neighborhood/county?
5. Lastly, Is 2022 a good year to buy a house?

The aim of the project is to provide the best counties/areas in the USA to invest in for a national real estate developer, individual buyers, banks looking for a place to develop a new apartment building or to purchase. Another goal is to predict the house prices in a county in the next few months.

We are trying to predict the house prices using Machine learning algorithms XGBoost and Linear Regression considering factors such as Median income in a county, Crime rate in that county, public schools, hospitals, hospital ratings and unemployment rate in the county.

Below is the scope of the effort:

• Data Selection

• Identify relevant attributes

• Data Preparation

• Feature Selection/Feature Engineering

• Exploratory Data Analysis

• Data Visualization

• Model Selection

• Model Evaluation

• Expected Results / Outcome

Data sources or plan for data

We will use many different data sources for Historical Sale price, Demographic data, Mortgage Interest rate, Crime Data, Schools data.

The related datasets include:

Zillow Housing Price data

<https://www.kaggle.com/zillow/zecon>

DEMOGRAPHIC DATA

This dataset comprises demographic data like population, age, sex, race and income, published by the US Census Bureau.

<https://www.census.gov/programs-surveys/acs>

CRIME DATA  
This page contains the crime data by California cities, published by FBI.

<https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county>

PUBLIC SCHOOLS DATA  
The page contains all active, pending, closed, and merged public schools and districts, also contains their corresponding Zip code, published by California Department of Education.

<https://www.kaggle.com/carlosaguayo/usa-public-schools>

My Main data set contains the following –

* SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
* MSSubClass: The building class
* MSZoning: The general zoning classification
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access
* Alley: Type of alley access
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to main road or railroad
* Condition2: Proximity to main road or railroad (if a second is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Overall material and finish quality
* OverallCond: Overall condition rating
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Exterior material quality
* ExterCond: Present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Height of the basement
* BsmtCond: General condition of the basement
* BsmtExposure: Walkout or garden level basement walls
* BsmtFinType1: Quality of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Quality of second finished area (if present)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: Heating quality and condition
* CentralAir: Central air conditioning
* Electrical: Electrical system
* 1stFlrSF: First Floor square feet
* 2ndFlrSF: Second floor square feet
* LowQualFinSF: Low quality finished square feet (all floors)
* GrLivArea: Above grade (ground) living area square feet
* BsmtFullBath: Basement full bathrooms
* BsmtHalfBath: Basement half bathrooms
* FullBath: Full bathrooms above grade
* HalfBath: Half baths above grade
* Bedroom: Number of bedrooms above basement level
* Kitchen: Number of kitchens
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality rating
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: Garage condition
* PavedDrive: Paved driveway
* WoodDeckSF: Wood deck area in square feet
* OpenPorchSF: Open porch area in square feet
* EnclosedPorch: Enclosed porch area in square feet
* 3SsnPorch: Three season porch area in square feet
* ScreenPorch: Screen porch area in square feet
* PoolArea: Pool area in square feet
* PoolQC: Pool quality
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

Topics covered and questions to answer from the data:

* What is the median price in given zip/city – this should tell me which states are the best to buy home at the moment.
* Which are the prominent variables influencing the price in a given market
* How much does the demographic data influence the pricing
* Which is the most popular month to buy a house – meaning lower prices within a year
* How does school rating matter to the housing prices ?
* How many variables can we consider in our overall strategy of predicting the housing prices.

**Exploratory Data Analysis (EDA)**

Build EDA model to Visualize the dataset, performing data mining and cleaning before applying the advanced regression models for getting my Predictive Analysis outcome. Using Machine learning algorithms XGBoost and Linear Regression considering factors such as Median income in a county, Crime rate in that county, public schools, hospitals, hospital ratings and unemployment rate in the county

I would basically need to filter my data source elements, ignore attributes that may not be required, apply null values to missing fields, visualize data elements and just restrict my predictive Analysis theorem to few major cities or just one city with details PA analysis.

After Data Visualization and cleaning the data, I am planning to check for correlation to answer key questions above from EDA. Also, I will be removing unwanted data and outliers.

Once we have the data and target values in 2 different variables, we can divide the data into two parts: the testing data and training data. The theory behind dividing the dataset into two parts is to ensure the model doesn’t overfit the training data. Otherwise, the model will perform well on the training data and perform poorly on the test data.

1. Which states should you buy a house in or rent?

Chart, histogram

Description automatically generated

2. Home value change for top expensive states

Chart, line chart

Description automatically generated

3. Median Home value per square foot in different states

Home prices per square feet over the yearsChart

Description automatically generated with low confidence

4. Price per square feet in the last 2 decades.

Chart

Description automatically generated

1. Top 6 states with median price

Table

Description automatically generated

1. Bottom 6 states with median price per square feet

Background pattern

Description automatically generated with medium confidence

1. Median Listing price of all homes

Chart, histogram

Description automatically generated

**Model Creation :-**

Predictive Model with XGBoost and Linear Regression - For building predictive model of house prices with Zillow Economics dataset, I have used County\_time\_series data as it provides house prices for all homes in a county. Different factors considered for predicting the house prices are Median Income, Crime Rate, Public schools, Hospitals and Hospital ratings, Unemployment rate in that county.

As bagging algorithms, Random Forest and Extra Tree Classification were chosen, and XGBoost was picked as one of the boosting techniques. As one tree algorithm, the Decision Tree algorithm was used.

More on this once we execute the above methods and play around.

**Model Evaluation/Selection**

This is in progress.

**Summary/Conclusion**

This is in progress

**Reference :**

<https://www.kaggle.com/zillow/zecon>

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

<https://finance.yahoo.com/news/housing-index-shows-renting-beats-193900999.html>

<https://towardsdatascience.com/house-price-prediction-with-zillow-economics-dataset-18709abff896>

<https://www.section.io/engineering-education/house-price-prediction/>

<https://yalantis.com/blog/predictive-algorithm-for-house-price/>

<https://www.kaggle.com/erick5/predicting-house-prices-with-machine-learning/data>

<https://rpubs.com/nweissm/670674>