Udacity AWS Machine Learning Nanodegree

Capstone Proposal

Predicting House Prices with Machine Learning Algorithms

Pakiza Valizada

March 2023

1) Domain Background

Real estate market is considered one of the most volatile industries since house prices are very prone to change depending on economic conditions of the countries. According to Maslow's "Hierarchy of Needs", house is considered one of the most essential needs of humans. Therefore, the main aim of this project is to predict the house prices based on various features (size of the house, number of the rooms, number of the bathrooms, location, availability of garage and etc.)

The AWS SageMaker Studio (Python 3 - jupyter notebook) and "ml.m5.xlarge" instance type will be used in order to run the model.

2) Problem Statement

The main aim of my research project is to determine which variables are correlated and how these variables can be used to predict the house prices. Client house sellers want to buy a house at a reasonable price with the highest return, at the same time, house buyers want to make sure that they want to get a fair price on houses.

I'll use various regression techniques to determine the prices of houses based on their features.

3) Dataset and Inputs

Dataset and Inputs can be found on Kaggle's website:

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Dataset consists of the files below: [1]

- **train.csv** - the training set
- **test.csv** - the test set
- **data_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data description file consist of data fields below:

- **SalePrice** - the property's sale price in dollars. This is the target variable that we'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

The “SalesPrice” will be used as a dependent variable in the model. To determine the distribution of the labels in the dataset, I’ll use probability plot, joint plot, boxplot, barplot, and etc.

4) Solution Statement

The proposed solution to this problem is to apply various regression techniques, such as LASSO, ElasticNet, Gradient Boosting (GBM) to predict the house prices. Firstly, I’ll split the data into train and test sets, and extract Lasso, ElasticNet, Gradient Boosting scores. After that, I’ll use evaluation metric to evaluate the test data.

5) Benchmark Model

For this model, I’ll use the algorithms outlined in “Winky K.O. Ho, Bo-Sin Tang & Siu Wai Wong (2021) Predicting property prices with machine learning algorithms, Journal of Property Research.”[2]

6) Evaluation Metrics

The evaluation metric for this project will be Mean Absolute Error (MAE).

7) Project Design

Before building the regression models, we need to do preprocessing and feature engineering in order to improve the accuracy of our model.

Firstly, I’ll do preprocessing (imputation of null values and cleaning data) and feature engineering after identifying the various types of features. After finishing preprocessing and feature engineering, I’ll split the dataset into “train” and “test” sets (80-20).

References:

- https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data[1]
- Winky K.O. Ho, Bo-Sin Tang & Siu Wai Wong (2021) Predicting property prices with machine learning algorithms, Journal of Property Research, 38:1, 48-70, DOI: 10.1080/09599916.2020.1832558[2]