PROJECT REPORT ON HOUSING PRICE PREDICTION

Team Members

GANESH CHITLAPALLY, SAMEER SHAIK AND SRI RAM DEGALA

ABSTRACT

This research project dives into the field of machine learning, with a particular emphasis on property price prediction. The main goal is to use various strategies to reduce Mean Square Error (MSE) within a linear model framework. To enhance and supplement the dataset, the project includes extensive preprocessing, exploratory data analysis (EDA), and complicated feature engineering approaches. The study used a variety of algorithms, including Linear Regression, Ridge Regression, Random Forest, Support Vector Regression, Gradient Boosting, and XGB Regressor, to determine the most efficient approach, with the goal of optimizing the prediction model's overall efficiency. The ultimate objective is to discover insights that aid in the decrease of MSE, hence improving the accuracy of home price projections.

INTRODUCTION

The goal of this research is to use sophisticated machine learning techniques to create a forecast model for home prices. Property price forecast accuracy is crucial in the real estate industry. The goal of this research is to create a model that can forecast property prices using a large dataset that includes a wide variety of dwelling features. Predicting property values in the volatile real estate market necessitates a complex strategy. Harnessing the potential of machine learning algorithms, particularly in terms of minimizing Mean Square Error, offers the possibility of improving prediction accuracy. This introduction lays the groundwork for the investigation of approaches meant to negotiate the complex interactions between various elements impacting house values.

BACKGROUND

Navigating the complex environment of home price prediction necessitates a smart methodology that considers the various forces that govern the real estate market. A careful and strategic decision is taken when selecting a machine learning model to anchor the investigation inside the framework of the linear model. This choice is not random, but is motivated by the linear model's intrinsic interpretability, which provides a visible lens through which to examine the correlations between input variables and home prices. The linear model serves as the foundation for a thorough investigation of forecast accuracy. The focus here is on minimizing Mean Square Error (MSE), an intentional emphasis aimed at improving forecast precision. MSE, as a statistic, correlates with the ultimate goal of developing not just predictive capabilities, but also minimizing variances between expected and actual values.

OBJECTIVE

The primary goal of this research is to create a trustworthy machine learning model capable of providing accurate price forecasts in the real estate industry. This model's application is intended to meet the different demands of potential buyers, sellers, and real estate analysts.

DATA PREPROCESSING

ABOUT DATA

The dataset used in this project has shape of (1022,82) which represents 1022 rows with 82 features. The dataset came in two pieces, and they are train and test data sets. As the data is already divided into train test datasets, it not required train_test_split in the project. Detailed attributes related to house specifications, conditions, and prices are 'Unnamed: 0', 'Id', 'MSSubClass', 'MSZoning', 'LotFrontage', 'LotArea', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'MasVnrArea', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinSF1','BsmtFinType2', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', 'Heating', 'HeatingQC', 'CentralAir', 'Electrical', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual', 'TotRmsAbvGrd', 'Functional', 'Fireplaces', 'FireplaceQu', 'GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageCars', 'GarageArea', 'GarageQual', 'GarageCond', 'PavedDrive', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'PoolQC', 'Fence', 'MiscFeature', 'MiscVal', 'MoSold', 'YrSold', 'SaleType', 'SaleCondition', 'SalePrice'.

The Target Column is SalePrice.

DATA CLEANING

Standardization of Column Names: Column names were standardized for consistency.

Handling Missing Values: Columns with a significant amount of missing data were identified and removed. The remaining missing values were imputed appropriately. The threshold for removing and imputing used was 50%. The columns with above 50% missing values were removed and for columns below 50%, simple imputer method is used. Imputations were made by central tendencies like mean, median and mode for respective column requirement.

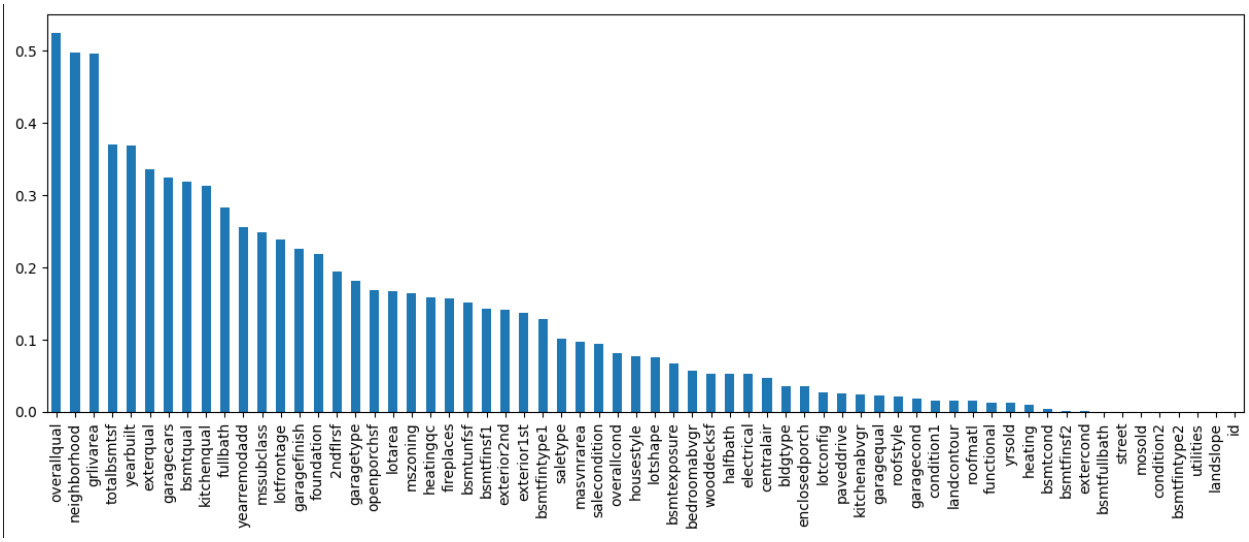
FEATURE ENGINEERING

Imbalanced Data: Removed columns with data imbalance over 90%. The data was 90% same throughout the datapoints. The columns can be considered as constant columns as the variance of these columns is approximately 0.

Correlation Analysis: Highly correlated features were identified and removed to reduce multicollinearity.

Feature Selection: Features with the highest importance scores, as determined by mutual information regression, were selected for model training. Bar Plot was plotted for the feature importance and by using percentile selection, selected 55% of the important columns which summed to 35 columns excluding target column.

The Final Columns after preprocessing and Feature selection is 'mssubclass', 'mszoning', 'lotfrontage', 'lotarea', 'lotshape', 'neighborhood', 'housestyle', 'overallqual', 'overallcond', 'yearbuilt', 'yearremodadd', 'exterior1st', 'exterior2nd', 'masvnrarea', 'exterqual', 'foundation', 'bsmtqual', 'bsmtexposure', 'bsmtfintype1', 'bsmtfinsf1', 'bsmtunfsf', 'totalbsmtsf', 'heatingqc', '2ndflrsf', 'grlivarea', 'fullbath', 'halfbath', 'kitchenqual', 'fireplaces', 'garagetype', 'garagefinish', 'garagecars', 'openporchsf', 'saletype', 'salecondition'.



MODEL DEVELOPMENT

SUBGROUPING

Kmeans Clustering: The dataset was divided into subgroups using KMeans clustering based on the features. The K value considered for clustering is 3. The optimal K value is considered after testing various K values. After clustering the training and test the data distributions into 3 groups (0,1,2) are

Train data – {(0, 58), (1, 634), (2, 330)}

Test data – {(0, 17), (1, 290), (2, 131)}

MODEL TRAINING

Algorithms Used: The algorithms which are applied in this model are LinearRegression, Ridge, SVR, RandomForest.

Hyperparameter Tuning: GridSearchCV was used for hyperparameter tuning and validation of each model.

METRICS

Mean Squared Error (MSE), R2 Score, and Mean Absolute Error Percentage(MAPE) were used to evaluate the models' performance.

STACKING METHOD

Model Integration: A stacking approach was used to combine predictions from individual models. RandomForestRegressor served as the final estimator.

The results obtained after the algorithms implementations are.

```
MSE for Group 0: 97818351.73267063
MSE for Group 1: 112597.77499574132
MSE for Group 2: 768463.1488796958
And the Average MSE value is 32899804.21884869.
```

For further Improving the MSE value we used **Random Forest, Gradient Boost Regressor and XGB Regressor**. And by using these Algorithms we were able to bring down the MSE value to 9881 With 99% accuracy. The best model which performed well on this data is Gradient Boost Algorithm.

MODEL EVALUATION AND RESULTS

The performance of each model in each subgroup was reported in the code file.

The average MSE across all groups and the best-performing model within the stacking method were identified.

Model: Random Forest

```
  Group 0: MSE: 87760912.20653531, R2: 0.9936455584156629, MAPE: 0.0132721
33120136798
  Group 1: MSE: 132180.355065694, R2: 0.999909481200919, MAPE: 0.000884781
190770835
  Group 2: MSE: 1042358.5315331412, R2: 0.9996494874321803, MAPE: 0.000982
7307919162206
```

Average MSE for Random Forest: 29645150.364378046

Model: Gradient Boosting

```
  Group 0: MSE: 1386.2393882185256, R2: 0.9999998996275563, MAPE: 6.428856
385624887e-05
  Group 1: MSE: 6199.755609708676, R2: 0.9999957543280005, MAPE: 0.0003773
977773399077
  Group 2: MSE: 22056.68011424603, R2: 0.9999925830284393, MAPE: 0.0004989
552621241099
```

Average MSE for Gradient Boosting: 9880.891704057745

Model: XGBRegressor

```
  Group 0: MSE: 71111.72274464574, R2: 0.9999948510643659, MAPE: 0.0002582
9758639394045
  Group 1: MSE: 138093.83341727272, R2: 0.9999054315752502, MAPE: 0.001674
6796673866942
  Group 2: MSE: 36131.50983646277, R2: 0.9999878501034828, MAPE: 0.0006358
826746792291
```

Average MSE for XGBRegressor: 81779.0219994604

The best-performing model overall is Gradient Boosting with an average MSE of 9880.891704057745.

CONCLUSION:

The project successfully developed a sophisticated predictive model for housing prices, with Gradient Boosting standing out as the most effective algorithm. This model, characterized by its accuracy and reliability, can serve as a valuable tool for various stakeholders in the real estate market, aiding in informed decision-making and market analysis. The use of advanced machine learning techniques and the implementation of a stacking approach significantly contributed to the model's high predictive performance.