HOUSE PRICE PREDICTION USING MACHINE LEARNING

PHASE 1: PROBLEM DEFINITION AND DESIGN THINKING

- Developing an accurate prediction model for housing prices is always needed for socio-economic development and well-being of citizens.

- In this paper, a diverse set of machine learning algorithms such as XGBoost, CatBoost, Random Forest, Lasso, Voting Regressor, and others, are being employed to predict the housing prices using public available datasets.
- The housing datasets of 62,723 records from January 2015 to November 2019 is obtained from the Florida's Volusia County Property Appraiser website.

- The records are publicly available and include the real estate/economic database, maps, and other associated information.

- The database is usually updated weekly according to the State of Florida regulations. Then, the housing price prediction models using machine learning techniques are developed and their regression model performances are compared.

- Finally, an improved housing price prediction model for assisting the housing market is proposed. Particularly, a house seller/buyer or a real estate broker can get insight in making better-informed decisions considering the housing price prediction.

- The empirical results illustrate that based on prediction model performance, Coefficient of Determination (R2), Mean Square Error (MSE), Mean Absolute Error (MAE), and computational time, the XGBoost algorithm performs superior than the other models to predict the housing price.

TECHNIQUES:

Housing Price Prediction(Machine Learning Algorithms, XGBoost Method,Target Binning)

Case Study and Modeling Framework

- In this section, a general overview of the case study and a real-world case of house pricing problem is presented.

- In addition, this section includes the information coming from the property sales datasets of the Volusia County Property Appraiser and the proposed modeling framework to analyse the datasets.

Problem Description and Research Framework:

- The primary aim of this case study is to predict housing price for the given features to maximize the prediction accuracy by employing the proposed methodology.

- This housing problem can be considered both as a regression or a classification housing problem.

- Since the classification problem was previously reported in the literature, this research considers several regression models with target variable binning which are applied on the housing market data to predict the property price.

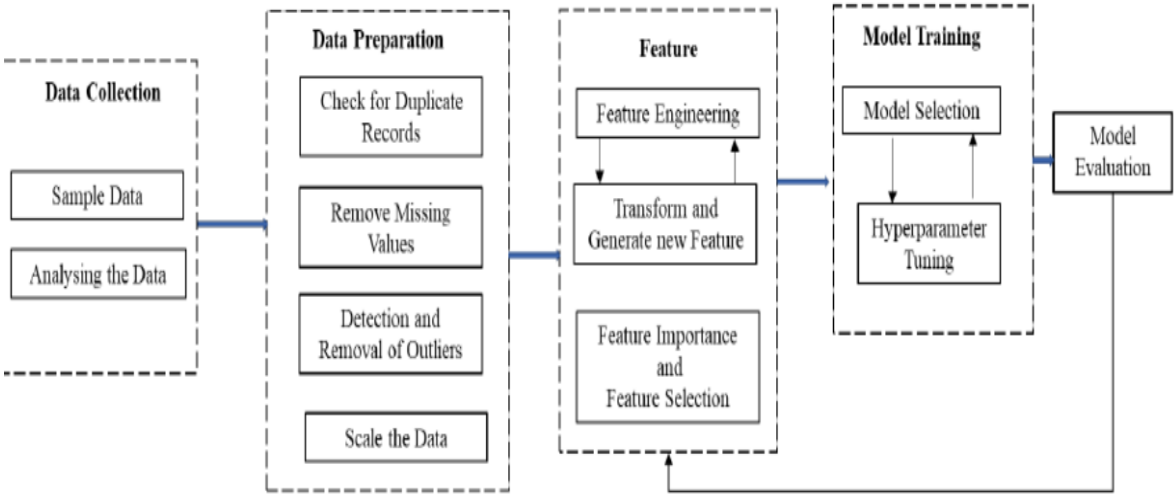

Fig. 1. Research framework for the housing price problem

Removing Duplications:

- For regression modeling, if some data points are replicated by being present more than once in the dataset, they are more strongly represented in the underlying data, so the regression algorithm treats them as having more importance.
- It can be thought of each occurrence of a data point as pulling the regression line towards it with the same force.
- If there are two data points at a given point in the regression model plane, they will pull the line towards them twice as hard.
- Therefore, it is indicated to remove the duplicate values from data itself.

- Duplication removal should be done carefully though, as these duplicate data points may cause what is called data leakage, when splitting the entire data into training and validation sets.

- If a data point in the training set has a duplicate value in the validation set, then the model will give biased prediction toward these duplicate data points, which is not desirable since will bias the entire prediction model.

- The duplicate entries and null entries from the housing sale transactions dataset were removed.

Feature Engineering:

- The main purpose of feature engineering is to find the most influential or partially important features of the dataset and detect the less valuable features to be removed from the dataset.

- The process results in a highly efficient and less complex model.

- Following this process requires domain expert knowledge in identifying a set of key features of the dataset and performing feature analysis, which is described in the next subsection.

Feature Analysis:

- The main purpose of feature engineering is to find the most influential or partially important features of the dataset and detect the less valuable features to be removed from the dataset.

- The process results in a highly efficient and less complex model.

- Following this process requires domain expert knowledge in identifying a set of key features of the dataset and performing feature analysis, which is described in the next subsection.

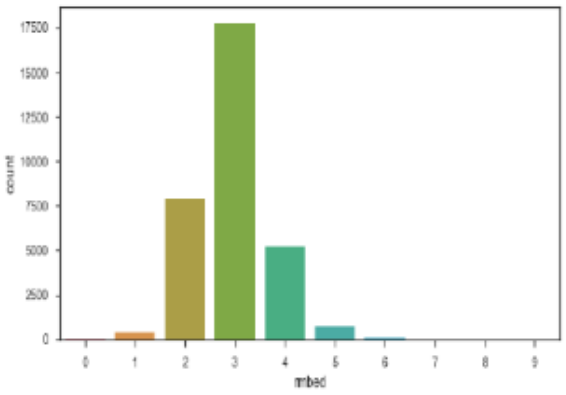

Fig. 6. Number of bedrooms per house

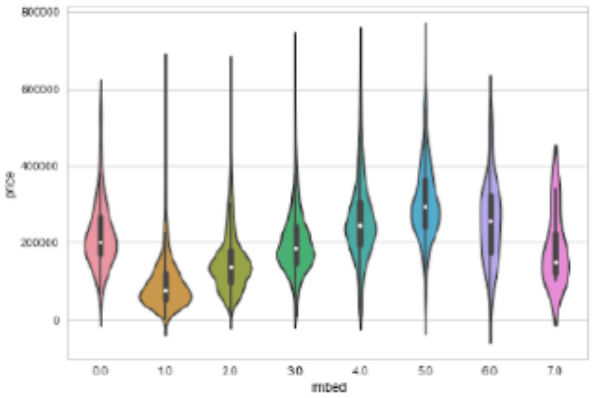

Fig. 7. Violin Plot between bedrooms and property sale price

Random Forest Method:

- Random forest (Breiman, 2001) is an ensemble of decision trees where each tree is built from a sample drawn from the training set.
- To give more randomness in building a random forest, some random subset of given features or all features are considered for best split, while splitting operations on each node (Ho, 1998).
- Size of the random subset is passed by the user as a hyper parameter.
- The individual decision tree suffers from high variance problems that lead to overfitting of the tree estimator.
- Random forest overcomes the problem of high variance in individual tree by providing above-mentioned two types of randomness.
- Random subset samples make different errors, and thus estimators generalize well by taking the uniform average of each predictor that helps in cancelling out the errors.
- Generally, random forests suffer from the increased bias problem, but variance is the key point to take care over bias.

XGBoost Method:

- XGBoost (Chen & Guestrin, 2016) stands for "Extreme Gradient Boosting" and is a technique based on the concept of gradient boosting trees (Friedman, 2001).

- The main difference from other gradient boosting based techniques is the objective function, which consists of two parts: training loss and regularization term, as presented in the equation (5).

$$\mathcal{L}(\emptyset) = \sum_i \ell\left(\hat{y}_i, y_i\right) + \sum_k \Omega(f_k) \qquad (5)$$

$$\text{where, } \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$$

- The training loss measures how predictive the model is with respect to the training data.
- The regularization term controls the complexity of the model that helps to improve the model generalization.
- A common choice of training loss is the mean squared error. In case of XGBoost, the Taylor expansion of the loss function up to the second order is used to expand the polynomials loss function.

Conclusions :

- This study employs machine learning techniques, with and without target binning, to develop a price prediction model for housing problems.

- It uses a rather large publicly available dataset of real estate transactions for a 5-year period.

- The regression model performances of the models are compared with one another and with the benchmark model.

- The empirical results show that the XGBoost algorithm with target binning provides superior performance for all metrics under study the coefficient of determination R^2 score, the mean errors, and the computational time.
- The developed model may facilitate the prediction of future housing prices and the establishment of policies for the real estate market.
- Particularly, the sellers and buyers of properties can benefit from this study and make better-informed decisions regarding the

property evaluation.

- In addition, property agents can focus on the seasonality effects, especially during the summer season, when most of the people buy their properties, and on the clear preference for two- or three-bedroom properties.

- The financial organizations and mortgage lenders may also find the study beneficial and identify more accurate real estate property value, risk analysis, and lending decisions.

- The study can be enlarged in a subsequent research by increasing the dataset size so potentially uncovered details and features of the dataset and of this study can be addressed.

- An increased dataset would potentially be good enough for employing deep neural networks, which can assure that more in-depth analysis on the housing price prediction can be performed.
- Then, the enlarged housing price prediction problem can be tackled as a classification problem.