**Project report for ESI6612**

Statistical Data Intelligence

Fall 2023

University of South Florida, Tampa



**House Prices-Advanced Regression Techniques**

**Presented by**

Lohidhar Kolla- U08206440

Srivalli Chintapalli- U26366427

Santhosh Kunooru- U57217608

**Master of Science in Industrial Engineering**

**Under the guidance of**

Dr. Walter Silva Sotillo

Mr. Kim Huynh

**Department of Industrial and Management Systems Engineering**

**Contents**

**Introduction**

**1.1. Background Information**

Data analytics in predicting housing prices empowers real estate agents, buyers, and analysts in making data-driven decisions about the sale price, buying trend, future market predictions, and investment profitability. Traditionally, appraisers estimate the worth of real estate based on explicit market data, ultimately judging prices based on their personal opinions. As such, human intervention will unavoidably impact the evaluated price. However, because of the disparity in information in house markets, people frequently lose money on their investments. With recent advancements in machine learning, a subset of artificial intelligence (AI) used to assess and forecast outcomes, and growth in the collecting and utilization of large data, several industries, including real estate, have opted for them to improve decision-making.

**1.2 Problem Statement**

For this project, we use the dataset with 79 explanatory variables describing almost every aspect of the residential homes in Ames, Iowa(Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle). The task at hand is to employ advanced regression techniques to help us understand the relationship between the various features of the house, and the relation between these features, and build models to predict the sale price.

* 1. **Objectives**

The key objectives of this project are to have accurate predictions, use feature engineering to identify and incorporate key features, preprocess the data to address missing values and potential outliers and cross-validate the models to assess model generalizability.

* 1. **Scope**

This project's scope is limited to using the features in the supplied dataset to forecast housing prices. It doesn't account for factors that could affect the housing market, such as macroeconomic indicators, inflation, and geopolitical situations.

**Literature Review**

One of the necessities for human existence, along with food, water, and many other things, is a place to live. As people's living standards increased over time, there was a sharp increase in the demand for homes. Although some individuals purchase a home as an investment or piece of real estate, most people worldwide purchase a home for habitation or employment. Each year, there is an increase in the demand for homes, which subsequently drives up home prices annually. House price prediction can be done by using multiple prediction models (Machine Learning Model) such as support vector regression, artificial neural network, and more [1]. Machine Learning alogirithms can analyze historical data and other relevant factors such as demographics,location,size, and amenities to aacurately predict the value of a property[2]. Additionally, it enables consequential learning and enhances model predictions through the methodical addition of more recent data[3]. Three broad categories can be used to classify machine learning: semi-supervised learning, unsupervised learning, and supervised learning. Essentially, the goal of supervised machine learning algorithms is to find a function that can provide reliable out-of-sample predictions.

For example, in property research, if an investigator intends to make forecast of housing prices *yi* from its physical, neighbourhood and accessibility characteristics *xij* from a sample of *n* apartments, one can assume L(*yi*ˆ,*yi*) to be the prediction loss function. A machine learning algorithm will look for a function *f*ˆ that produces lowest expected prediction loss *E*(*yi*,*xij*)[L(*f*ˆ(*xij*),*yi*)] on the *test* data from the same distribution[4]. Support vector machines and linear regression are two examples of supervised learning. Unsupervised learning in ML learns from the data without human supervision, . Semi supervised machine learning handles a mixture of both labeled and unlabled data. For this project we chose advanced ML techniques like Random forest [5], Extratrees regressor, XGBoost and ensemble model as they have shown to provide accurate predictions[6].The methodology of this paper is comprised of the following steps.

1. Data Mining, Data Cleaning and Preprocessing
2. visualizations and Feature Selection
3. Model Training and Validation
4. Hyperparameter Tuning and ensemble models.

The data from Kaggle consists of pre-split data sets, named “Train” and “Test”

A computer code with numbers and letters

Description automatically generated

A black text on a white background

Description automatically generated

* The test data set consists of 1459 observations with 80 variables.

The preprocessing stage of our analysis was critical in ensuring the quality and usability of the dataset. After importing the training and testing data from the Kaggle competition, we embarked on a thorough cleaning process. We began by identifying features with a significant number of missing values. Any feature with more than 70% missing data was considered unreliable for our predictive models and was subsequently removed. This rigorous approach to feature selection helped to maintain the robustness of our models. For the remaining features, missing values were meticulously handled to preserve the integrity of the dataset. We imputed missing values in categorical features with the most frequent value, or mode, and filled missing numerical values with the mean of the respective feature. This strategy was chosen to mitigate the impact of missing data on our predictive analysis.

**Log Transformation:**

A math equation with a number and a square

Description automatically generated with medium confidenceValidation

**A graph with numbers and a number of lines

Description automatically generated with medium confidence**