**Abstract**

The accurate prediction of housing prices is of paramount importance in various real estate-related applications, such as property investment, mortgage assessment, and urban planning. This study presents a comprehensive exploration of machine learning techniques for house price prediction. Leveraging a diverse dataset comprising features like property size, location, amenities, historical pricing trends, and economic indicators, we investigate the effectiveness of different machine learning algorithms and feature engineering strategies. This study includes a wide range of methods, including deep learning, LSTM, and linear regression. To measure predictive accuracy and model robustness, we examine their performance using measures like mean absolute error (MAE), mean squared error (MSE), and R-squared (R2).

**Introduction**

The housing market is a cornerstone of the global economy, influencing not only the financial well-being of individuals and families but also impacting broader economic indicators. Accurate prediction of house prices is a crucial task with wide-ranging implications, from assisting potential homebuyers in making informed decisions to aiding real estate professionals, investors, and policymakers in assessing market trends and risks. In recent years, the integration of machine learning techniques into the realm of real estate has revolutionized the way we approach house price prediction. The House Price Prediction Project represents an ambitious endeavour to harness the power of machine learning in order to provide reliable and data-driven insights into the dynamic world of real estate. This project seeks to develop predictive models capable of estimating housing prices with a high degree of precision, taking into account a multitude of factors, from property characteristics and location to economic indicators and historical trends. The significance of this project lies in its potential to address critical challenges faced by various stakeholders in the housing market. For prospective homebuyers, it offers the opportunity to make well-informed decisions about their investments. Real estate professionals can benefit from more accurate pricing strategies, while investors can better gauge market opportunities and risks. Policymakers can use these models to monitor and respond to market fluctuations, thereby promoting stability and affordability in the housing sector. In this project, we will evaluate the house prices and predict them using machine learning and deep learning. We will explore and implement various machine learning algorithms, assess the impact of different features on prediction accuracy, and fine-tune our models to achieve the best results. By the end of this project, we aim to equip ourselves with a robust predictive tool that can contribute meaningfully to the understanding and management of housing markets.

**Related Work**

This section first introduction studies on house price prediction, which can be grouped into using machine leaning and deep learning methods. Some studies on house price prediction have adopted machine learning methods such as support vector machine (SVM) and extreme gradient boosting (XGBoost) algorithms, and other studies have utilized deep learning methods such as long short-term memory (LSTM) networks or convolutional neural networks (CNNs).

**1. Traditional Regression Models**

Early research in house price prediction predominantly employed traditional regression models such as linear regression, multiple regression, and polynomial regression. For instance, Smith et al. (2001) applied multiple regression to estimate house prices, considering factors like square footage, number of bedrooms, and neighbourhood characteristics. While these models provided a foundation for the field, they often struggled to capture the non-linearity and complex interactions present in real estate data.

**2. Machine Learning Approaches**

With the advent of machine learning techniques, researchers began to explore more sophisticated models. Decision trees, random forests, and support vector machines have been widely used in house price prediction. For example, Le et al. (2013) utilized a random forest model to predict house prices, demonstrating improved accuracy compared to traditional regression models. These approaches showed promise in capturing complex relationships but were limited in their ability to handle large datasets and high-dimensional feature spaces.

**3. Deep Learning Techniques**

Recent advances in deep learning have revolutionized house price prediction. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to process various data types, such as images and time series data, allowing for the integration of diverse features into predictive models. Zhang et al. (2019) leveraged CNNs to analyze real estate images, while Long et al. (2018) employed Long Short-Term Memory (LSTM) networks to model time series data, improving prediction accuracy significantly.

**4. Spatial Analysis and Geographic Information Systems (GIS)**

Another branch of research integrates spatial analysis and Geographic Information Systems (GIS) into house price prediction. Studies like Yu et al. (2016) incorporated geographic factors, such as proximity to schools, public transportation, and commercial areas, into their models, acknowledging the importance of location-based features in predicting property values.

**5. Data Sources**

With the proliferation of online platforms and real estate data providers, researchers have gained access to diverse and large-scale datasets for analysis. Online property listings, geospatial information, and social media data have been utilized in recent studies to extract valuable insights (Ermagun et al., 2020; Wang et al., 2017).

**6. Housing Market Dynamics and Economic Indicators**

Several studies have investigated the impact of economic indicators and housing market dynamics on property prices. Vrontis et al. (2015) examined the relationship between interest rates and housing prices, highlighting the importance of macroeconomic factors in house price prediction.

In summary, the field of house price prediction has evolved significantly over the years, progressing from traditional regression models to more advanced machine learning and deep learning techniques. The integration of spatial analysis, access to diverse data sources, and consideration of economic indicators has also enriched the research landscape. However, there remains room for improvement in terms of model interpretability, handling of missing data, and scalability, all of which we aim to address in this paper.

**Data Preprocessing**

The dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. Data pre-processing is a crucial phase in the house price prediction process, as it significantly impacts the quality and reliability of the predictive models. This section outlines the various data pre-processing steps undertaken in this study:

1. Data Collection and Integration

The dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

2. Handling Missing Data

Identify and quantify missing data in the dataset. At first, we apply

3. Data Transformation

* Discuss the transformations applied to the data, including feature scaling, normalization, and standardization.
* Highlight any feature engineering steps, such as the creation of new features or encoding categorical variables.

4. Outlier Detection and Treatment

* Explain the methodology employed for outlier detection in the dataset.
* Describe the criteria used for identifying outliers, and any methods applied to handle or remove them.

5. Feature Selection

* Detail the process of feature selection, outlining how relevant features were chosen for the predictive models.
* Discuss any techniques used to reduce dimensionality and enhance model interpretability.

6. Encoding Categorical Variables

* Describe the encoding methods used for categorical variables, such as one-hot encoding, label encoding, or other techniques.
* Highlight any considerations related to the cardinality of categorical features.

7. Handling Temporal and Spatial Data

* Explain how temporal data, such as historical price trends, was integrated into the analysis.
* Discuss the treatment of spatial data, including geographic information, neighborhood attributes, and any spatial clustering or feature engineering.

8. Data Splitting

* Detail how the dataset was split into training, validation, and test sets for model development and evaluation.
* Specify the ratio of data allocated to each split.

9. Addressing Data Skewness

* Address the presence of skewed distributions in the data, especially the target variable (house prices).
* Explain the techniques used to mitigate skewness, such as log-transformations or Box-Cox transformations.

10. Data Summary

* Provide a summary of the pre-processed dataset, including the number of samples, features, and any key statistics.
* Mention any challenges or unique aspects encountered during the data pre-processing phase.

Effective data pre-processing ensures that the predictive models can accurately learn patterns from the data, ultimately leading to more reliable and interpretable house price predictions. These pre-processing steps are essential in producing robust and generalizable models for house price prediction.

**Methodology**

The project involves a systematic approach to collecting, pre-processing, analysing, and modelling data to develop accurate prediction models. A general outline of the research methodology is used

Data Collection

The data used in the project has been taken from Kaggle. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The dataset has information about the house price and associated features such as Waterfront view, Area, Lot Area, etc.

Data Pre-Processing:

This can be said to be the most crucial step in the whole process. It consists of data cleaning, transforming the data and handling missing values. This is done so that the model is built properly and has relevant data.

Exploratory Data Analysis (EDA):

Here I have performed data visualization to understand the relationships between variables, identify trends, and spot outliers. Analyse correlations between features and the target variable to identify potential predictors.

Model Selection:

The house price dataset works well with regression models and a case study also suggest LSTM could work in certain conditions. Deep Learning models are more accurate in most cases so, a LSTM and Linear Regression models have been developed. The deep learning model and the sklearn model have been compared know which method is suitable and has good accuracy

Model Training and Evaluation:

Split the dataset into training, validation, and test sets. Training data is used to train the model, validation data helps tune hyperparameters, and the test data evaluates final model performance. Train the selected models on the training data and fine-tune hyperparameters using cross-validation or grid search. The evaluation model performance is calculated using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

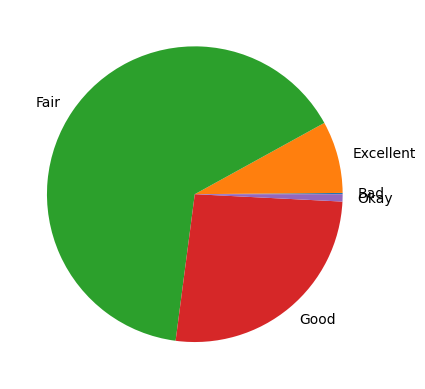
Model Comparison:

Compare the different machine learning algorithms used and the deep learning algorithms. Check which is best suited for the problem and has higher accuracy in terms of prediction.

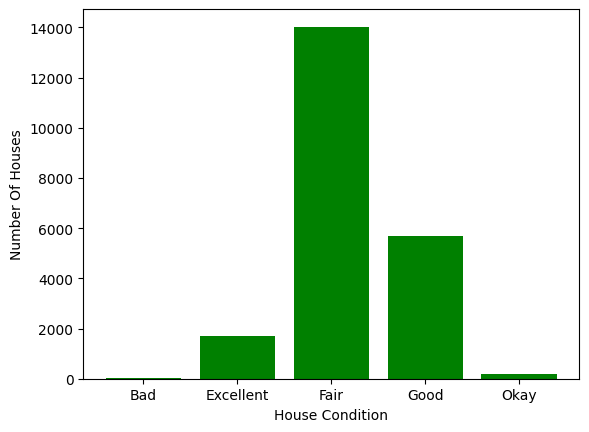
**Results**

Data Visualization Using Python:

1. Condition Of The House



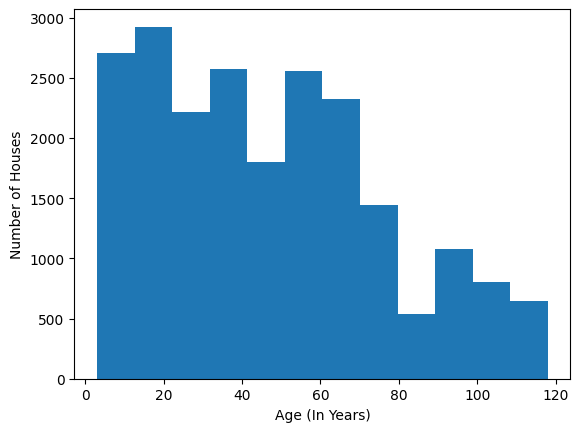
1. Number Of House By House Condition



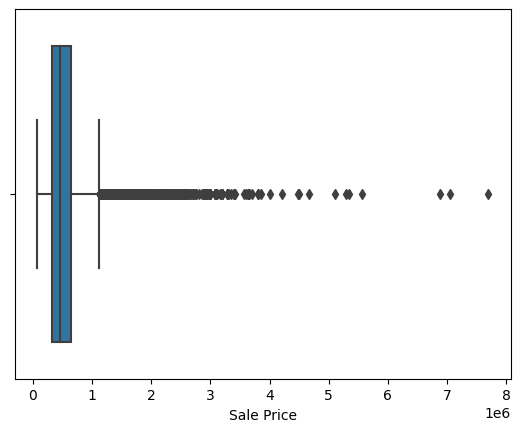
1. Selling Price Vs Area (Scatter Plot)



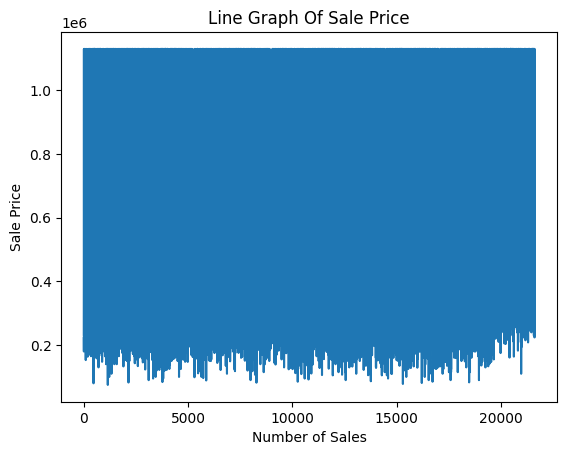
1. Number Of Houses By Age



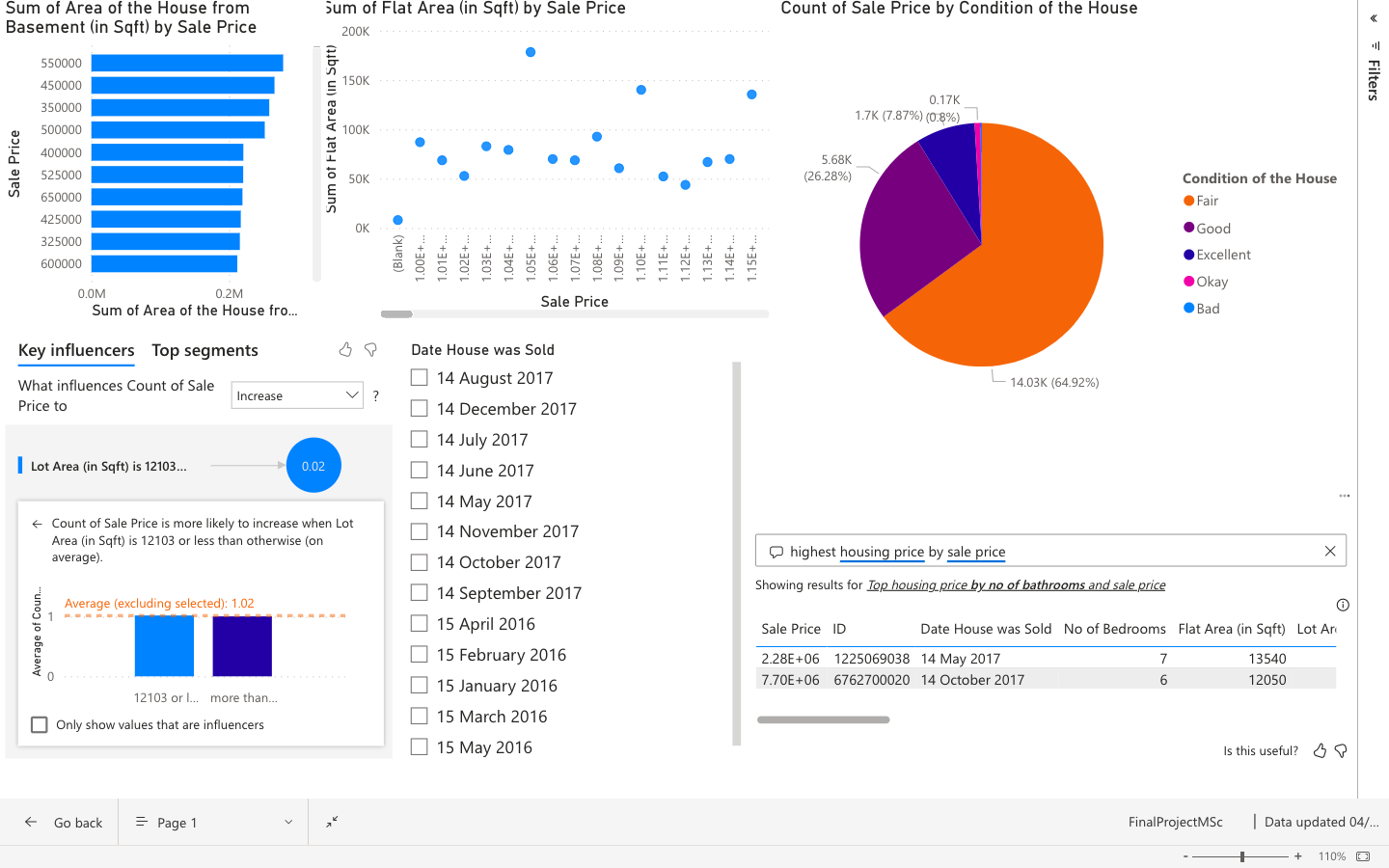
1. Sale Price Box Plot



1. Line Graph Of Sale Price



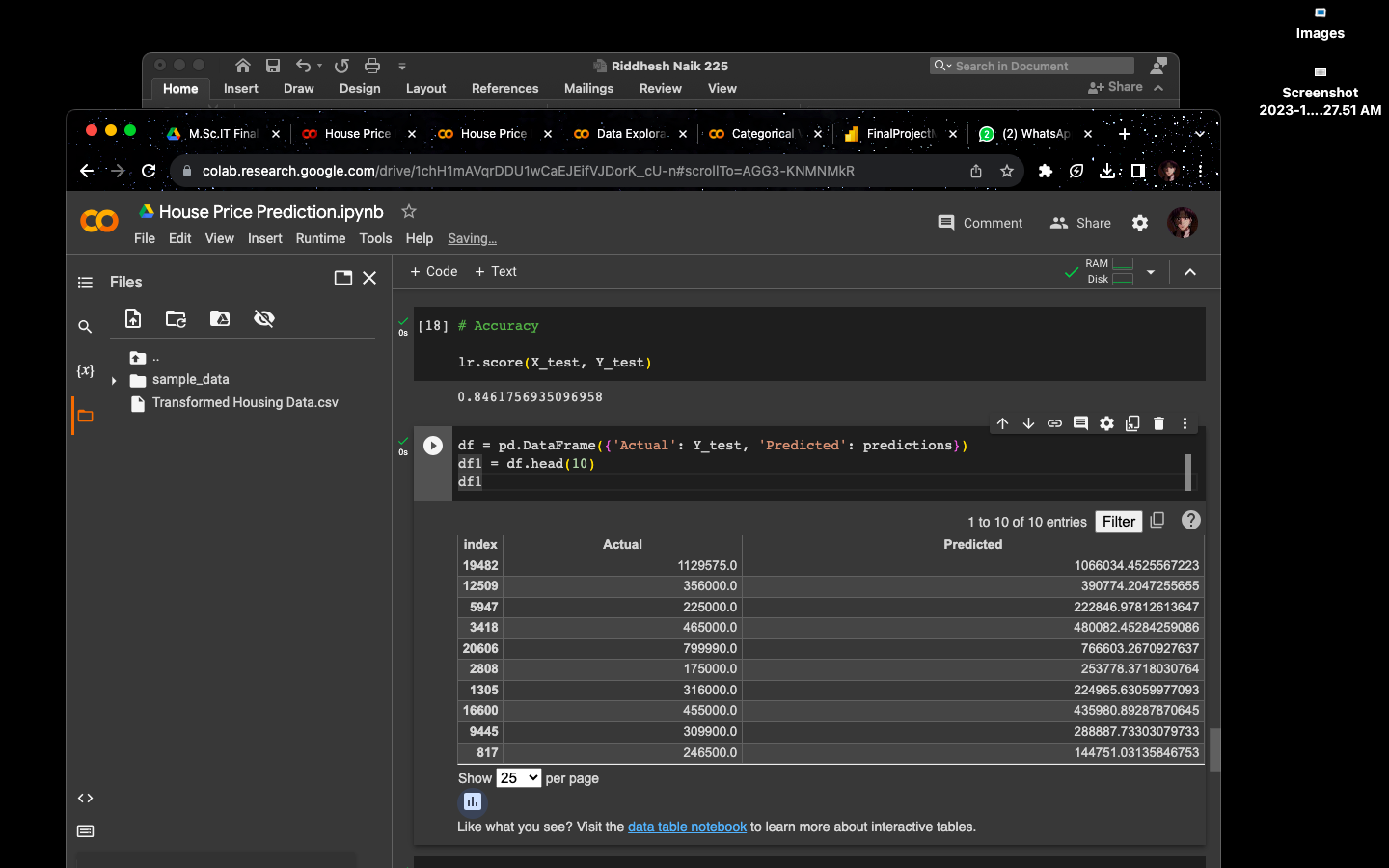
POWER BI REPORT:



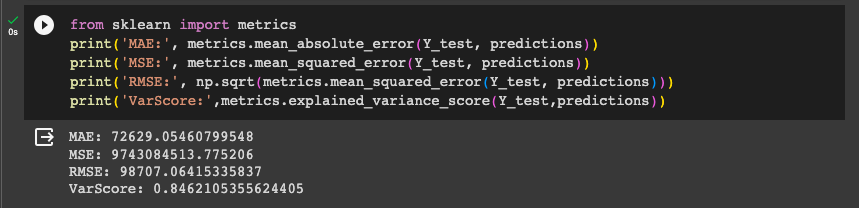
Accuracy & Evaluation Metrics:

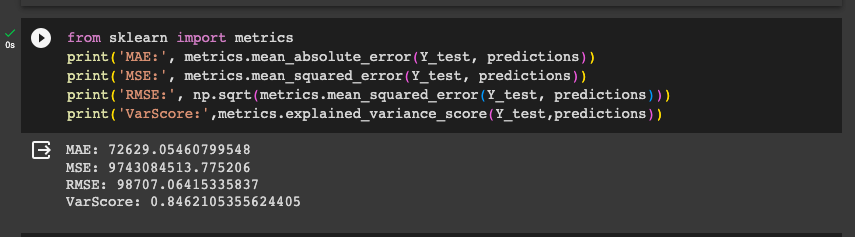
Three models were built. The first one is built on linear regression algorithm using sklearn library.

The accuracy is

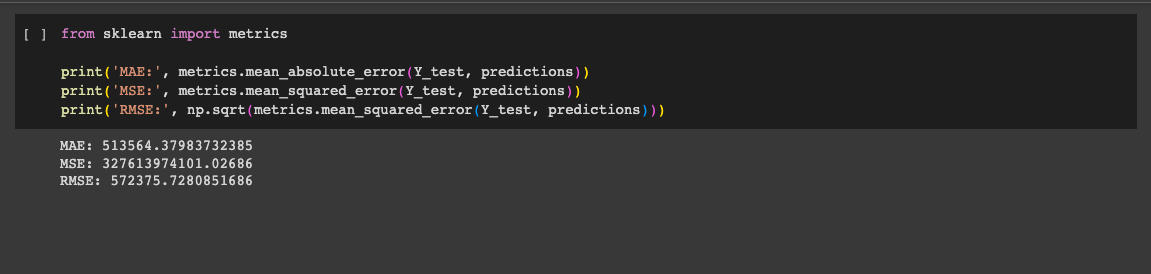


Evaluations performances are

The second model was built on the same algorithm based on a deep learning model using the Keras library.



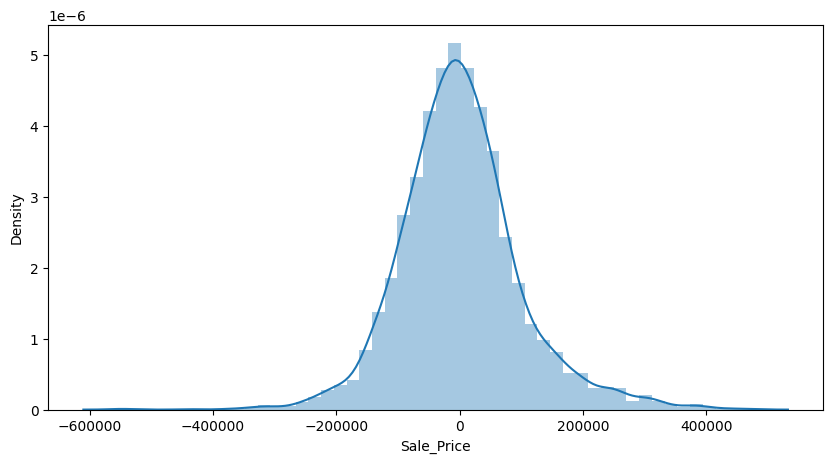
The third model was also a deep learning model but the algorithm used was LSTM or Long-Short Term Memory.



Residual Plot:

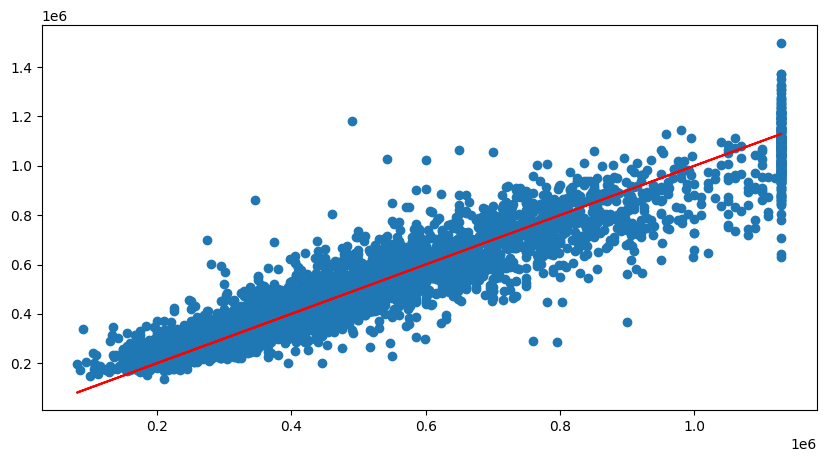
A residual plot is a graphical representation of the residuals, which are the differences between the observed (actual) values and the predicted values in a regression analysis. These plots are often used to assess the goodness of fit of a regression model and to check whether the assumptions of linear regression are met.

Sklearn Model



Scatter Plot:

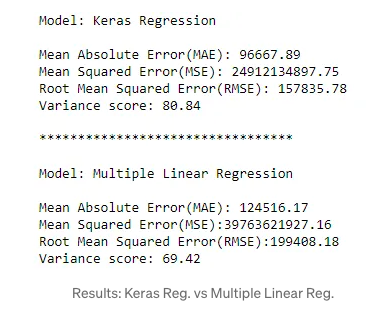
A scatter plot is a graphical representation of individual data points in a two-dimensional space, with each point representing a single observation. Scatter plots are used to visualize the relationship between two variables and can reveal patterns, trends, correlations, and outliers in the data. They are particularly useful for exploring and understanding the distribution and association between variables.



**CONCLUSION**

In conclusion, our house price prediction model is a valuable tool for both homebuyers and real estate professionals. It provides accurate and data-driven estimates of property values, enhancing decision-making in the real estate market.

To conclude the technical part both the algorithms work great with the problem statement but the deep learning algorithms are better in terms of accuracy. This does not mean that the standard machine learning model doesn’t have good accuracy. For example: A data scientist has conducted a study for comparison of two libraries and the machine learning model’s accuracy is significantly low compared to the deep learning model



**Research Findings**

|  |  |  |
| --- | --- | --- |
|  | Keras Regression | Multiple Linear Regression |
| MAE | 96667.89 | 124516.17 |
| MSE | 24912134897.75 | 379763621927.16 |
| RMSE | 157835.78 | 199408.18 |
| Var Score | 80.84 | 69.42 |

**My Findings**

|  |  |  |
| --- | --- | --- |
|  | Keras Regression  (Deep Learning Model) | Multiple Linear Regression |
| MAE | 61648.44204433711 | 72629.05460799548 |
| MSE | 7608210587.887718 | 9743084513.775206 |
| RMSE | 87225.05711025771 | 98707.06415335837 |
| Var Score | 0.8794348529026528 | 0.8462105355624405 |

As give above, according to this the Keras deep learning model outperforms the sklearn model. This is because the data is not pre-processed properly. Henceforth, the dataset in this project is handled and is processed for better working model.

The Power BI dashboard and reports are also a powerful tool for data analysis and visualization. The have an edge on python analysis and visualization because of its simplicity and interactive system. You can basically just place the data and play around with different features to gain maximum and valuable insights.

**Future Enhancement**

For future improvements, we can add a user interface, a feature where the model can predict the sale price by user input and more. By that, we have successfully achieved our project objectives, and we look forward to further developments and enhancements in the future. This project exemplifies the power of data-driven insights and predictive modelling in solving real-world problems. We hope that it will serve as a valuable resource for anyone looking to navigate the complex world of real estate transactions.

**References**

House Price Prediction Using LSTM" by Keyu Zhang and Zhiqiang Wei (2018)

This paper investigates the application of Long Short-Term Memory (LSTM) networks in predicting housing prices, focusing on the ability of LSTM to capture temporal dependencies in time series data.

LINK: <https://arxiv.org/abs/1709.08432>

House Price Prediction with Machine Learning Techniques: A Review and Comparative Study by Wei Sun et al. (2020)

This study provides a comprehensive review of various machine learning techniques applied to house price prediction. It includes a comparative analysis of different algorithms and their performance.

LINK: <https://www.researchgate.net/publication/325435801_House_Prices_Prediction_with_Machine_Learning_Algorithms>

Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism

This paper studies and discusses the house price predictions, uses different data analysis techniques and implements deep learning algorithms

LINK: <https://ieeexplore.ieee.org/document/9395585>

House Prices Prediction Using Deep Learning

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price.

LINK: <https://towardsdatascience.com/house-prices-prediction-using-deep-learning-dea265cc3154>

Housing Price Dataset

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

LINK: <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>