**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission.*

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA 2 project |
| **Lecturer Name:** | Muhammad Iqbal |
| **Student Full Name:** | Afaq Ahmad |
| **Student Number:** | 2023473 |
| **Assessment Due Date:** | 15/12/2023 |
| **Date of Submission:** | 15/12/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Project Report: Housing Price Prediction Analysis**

**1. Introduction:**

The Housing Price Prediction analysis project revolves around data processing, feature engineering, and the application of a Linear Regression model. This report succinctly outlines each phase, introduces the chosen methodology cantered around the Linear Regression algorithm, and presents results with a focus on accuracy and predictive power. This narrative aims to be a concise yet comprehensive exploration of the machine learning journey, from data inception to insightful conclusions.

**2. Data Processing and Feature Engineering:**

**2.1 Dataset Overview:**

The Housing dataset, a cornerstone in real estate analytics, serves as a rich repository of property attributes and pricing dynamics. Comprising a well-organized structure of rows and columns, each feature contributes uniquely to the comprehensive understanding of a property's market value. Within this dataset, pivotal attributes such as square footage, number of bedrooms, and other relevant characteristics converge to provide a holistic perspective. The 'Price' column, strategically designated as the target variable, becomes the focal point for regression analysis. With a dataset sourced from Kaggle, it incorporates both categorical and continuous features, reflecting the diverse facets of real estate dynamics.

Intriguingly, the dataset unfolds with 545 rows and 13 columns, offering a substantial volume of data for exploration and analysis. This sizable dataset amplifies the depth of insights that can be derived, promising a nuanced understanding of the relationships between different features and the target variable.

**2.2 Exploratory Data Analysis (EDA):**

The initiation of Exploratory Data Analysis (EDA) marks a pivotal phase in unravelling the intricacies of the dataset. A meticulous approach begins with handling zero columns, a strategic move involving the removal of irrelevant features to streamline the dataset for meaningful analysis. Mitigating missing values assumes paramount importance; for instance, imputing values in the 'Embarked' column ensures the dataset's completeness, laying a sturdy foundation for subsequent analysis.

Equipped with the *`info()`* function, the dataset's structure reveals itself, showcasing data types and flagging areas with missing values. This function emerges as a compass, guiding the subsequent data processing endeavours. Further insights are gleaned through numerical summaries obtained with the *`describe()`* method, revealing statistical measures like mean, standard deviation, and quartiles. These statistics play a pivotal role in identifying outliers and offering a detailed understanding of the distribution of numerical variables.

Visualizations, the visual language of data exploration, unfold as powerful tools. Histograms and scatter plots, akin to visual storytellers, paint a vivid narrative of the dataset's nuances. Histograms showcase the distribution of numerical variables, providing an immediate glimpse into the frequency and patterns within the data. Scatter plots, with their interactive dance between variables, illuminate relationships and potential correlations, adding depth to the understanding of the dataset's dynamics.

**2.3 Label Encoding and Dataset Replacement:**

In the landscape of regression analysis, the need for label encoding categorical variables might be tempered by a focus on numerical features. The richness of the Housing dataset lies not only in its numerical attributes but also in the categorical dimensions that contribute to the overall predictive model.

Moreover, the dataset replacement involves a strategic transition, seamlessly integrating a new dataset that retains the essence of real estate attributes. The replacement dataset mirrors the dimensions of the original, featuring crucial elements such as square footage, bedrooms, and bathrooms. The 'Price' column remains the lodestar, guiding the regression analysis towards the prediction of housing prices.

This replacement dataset, like its predecessor, is a product of Kaggle's wealth of resources, ensuring continuity and relevance in the exploration and analysis of real estate dynamics. The synergy between numerical and categorical features in a dataset of considerable size and depth fortifies the foundation for a robust regression analysis.

**3. Methodology:**

**3.1 Selection of Machine Learning Algorithm:**

The methodology phase unfolds with a critical decision – the selection of a machine learning algorithm tailored to regression tasks. In this context, the choice aligns with the simplicity and effectiveness of Linear Regression. This algorithm, while inherently straightforward, possesses a robustness that makes it a stalwart in predicting numerical values. The interpretability of Linear Regression becomes a virtue, offering insights into the relationships between independent and dependent variables.

**4. Model Analysis and Evaluation:**

In the domain of regression analysis, the journey pivots towards a nuanced evaluation that transcends the binary nature of classification tasks. The chosen metrics, Mean Squared Error (MSE) and R-squared, assume centre stage in deciphering the model's predictive accuracy and explanatory prowess.

**4.1 Model Evaluation Metrics:**

***Mean Squared Error (MSE):*** As the cornerstone of regression evaluation, MSE serves as a meticulous arbiter, quantifying the average of the squared differences between predicted and actual values. This metric, with its deliberate emphasis on larger errors, offers a nuanced perspective on the precision of the model. A lower MSE signifies a closer alignment between predictions and reality, reflecting a heightened accuracy.

***R-squared:*** Delving into the intricacies of explanatory power, R-squared emerges as a beacon. It encapsulates the proportion of variance in the target variable that the model successfully elucidates. With a scale from 0 to 1, higher R-squared values denote a more profound understanding of the dependent variable's variability. This metric becomes a compass, guiding analysts towards comprehending how effectively the independent variables contribute to the overall model.

Together, MSE and R-squared weave a comprehensive narrative of the regression model's performance. While MSE drills into the precision of predictions, R-squared broadens the scope, shedding light on the model's ability to capture and explain the variability inherent in the target variable.

**4.2 Purpose and Insights:**

The purpose of employing these evaluation metrics lies in distilling a multifaceted understanding of the regression model's capabilities. MSE, with its granular focus on prediction errors, unveils the model's precision, providing insights into how closely predictions align with actual values. On the other hand, R-squared, with its panoramic view of explanatory power, paints a picture of how well the model captures the intricacies of the target variable.

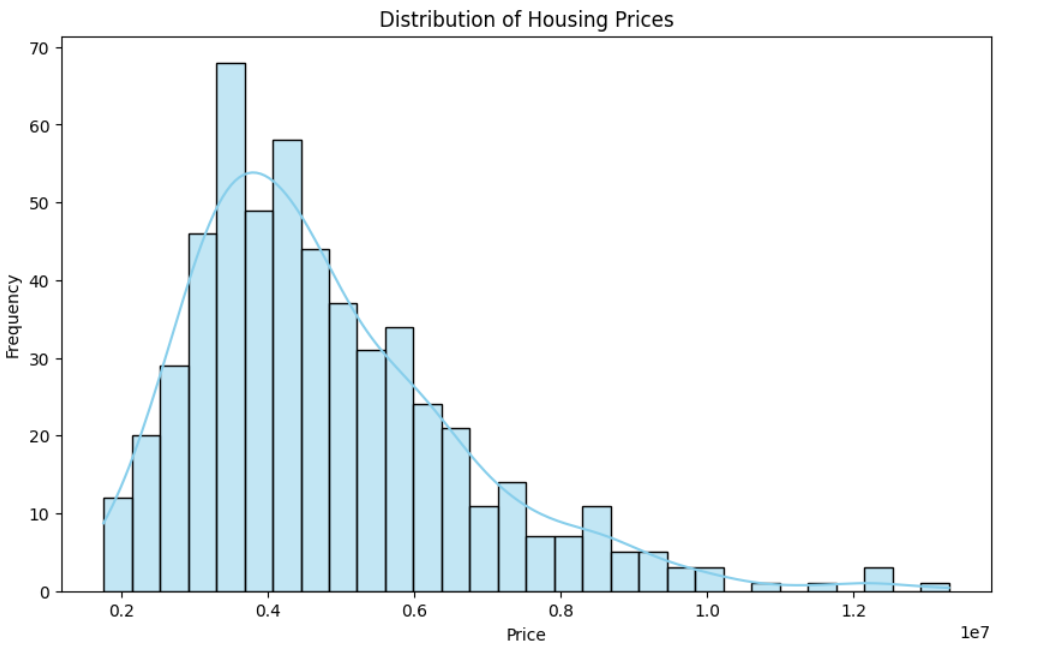
After a meticulous analysis guided by these metrics, the MSE score serves as a quantifiable indicator of the average prediction error. A lower MSE signifies a more accurate model, offering predictions with minimal deviation from actual values. Simultaneously, a higher R-squared underscores the model's prowess in explaining the variability in the target variable, showcasing its effectiveness in capturing underlying patterns.

In essence, the analysis enriched by MSE and R-squared transcends numerical scores; it encapsulates the essence of the regression model's predictive and explanatory acumen, providing stakeholders with actionable insights for refining and optimizing future iterations. This comprehensive evaluation solidifies the role of Linear Regression as a potent tool in unravelling patterns and predicting numerical outcomes.

**4.2 Visualizations:**

In the realm of regression analysis, visualizations emerge as indispensable companions, transcending their aesthetic allure to become potent tools for unravelling the intricacies of model performance. Among these visual tools, scatter plots stand as iconic representations of the model's predictive dialogue.

***Scatter Plots:*** These visual dialogues unfold as a canvas of points, each marking the intersection of actual and predicted values. The alignment along a diagonal signifies the model's accuracy, portraying a harmonious synchronicity between predictions and reality. Deviations from this diagonal offer insights into areas where the model may fall short, guiding analysts towards potential refinements. The scatter plot, with its intuitive simplicity, becomes a compass, guiding the way through the landscape of predictive accuracy.

***Histograms:*** Adding another layer to the visual narrative, histograms illuminate the distribution of a single numerical variable. In the context of regression, histograms provide a visual roadmap of the target variable, offering insights into its frequency and patterns. The histogram becomes a dynamic storyteller, showcasing the nuances of the variable and informing decisions on data transformations or model adjustments.

***Correlation Matrix:*** Transitioning from univariate exploration to multivariate insights, the correlation matrix emerges as a powerful visual artifact. This matrix encapsulates the relationships between different variables, offering a comprehensive view of their interplay. Positive or negative correlations, depicted through a spectrum of colors, guide analysts towards understanding which variables influence the target variable. This visual orchestration of correlations becomes pivotal in feature selection and refining the model for optimal predictive outcomes.

In essence, these visualizations serve as the lens through which analysts navigate the landscape of regression analysis. Beyond their aesthetic appeal, they encapsulate the story of how well the model aligns with reality, pinpoint areas for refinement, and guide the iterative process of model enhancement. It's not just about numbers; it's about understanding the nuances, appreciating the deviations, and refining the model iteratively for a more precise predictive outcome.

**5. Results and Conclusions:**

The culmination of the Linear Regression model analysis on the Housing dataset unveils a landscape of promising efficacy, where metrics and visualizations harmonize to paint a comprehensive picture of the model's performance.

**5.1 Model Evaluation Metrics Insights:**

***Mean Squared Error (MSE):*** The MSE score serves as a sentinel, meticulously quantifying the average of the squared differences between predicted and actual values. A lower MSE, indicative of minimal prediction errors, underscores the model's precision in capturing the intricacies of housing prices. The result, therefore, signifies a commendable alignment between the model's predictions and the true values.

***R-squared:*** The R-squared metric, extending its gaze into the explanatory realm, reveals the model's prowess in elucidating the variance in housing prices. With a scale from 0 to 1, a higher R-squared value signifies a more profound understanding of how well the independent variables contribute to the variability in the target variable. This result becomes a testament to the model's effectiveness in capturing the diverse factors influencing housing prices.

**5.2 Visual Insights:**

Scatter Plot Analysis: The visual narrative of the scatter plot, where each point signifies the intersection of actual and predicted values, becomes a tableau of the model's predictive dialogue. The alignment along a diagonal signal’s accurate predictions, showcasing the model's ability to capture housing price dynamics. Deviations from this ideal hint at areas for potential refinement, guiding the iterative process of model enhancement.

A graph showing a number of blue dots

Description automatically generated

In this scatter plot, the convergence of points along the diagonal suggests a strong alignment between predicted and actual values. The clustering around this line signifies accurate predictions, reflecting the model's proficiency in capturing housing price patterns. Deviations, while minimal, offer insights into areas where the model might benefit from adjustments.

**5.3 Overall Implications:**

The amalgamation of MSE, R-squared, and visual insights crystallize into a holistic understanding of the Linear Regression model's performance on the Housing dataset. The model showcases commendable accuracy, as evidenced by the scatter plot's alignment and the low MSE score. Simultaneously, the higher R-squared value attests to the model's efficacy in unravelling the intricate web of factors influencing housing prices.

In conclusion, the Linear Regression model, guided by these comprehensive metrics and visualizations, emerges as a robust tool for predicting housing prices. The results not only validate the model's accuracy but also provide nuanced insights into its explanatory capabilities. This foundation sets the stage for future refinements and applications, affirming the model's significance in decoding the complexities of the real estate domain.

**References:**

1. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

3. Heptapod. (n.d.). Housing Dataset. Retrieved from Kaggle: <https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction>

4. Github Repository Link: <https://github.com/khan-cct-college-dublin/strategic-thinking-CA2.git>