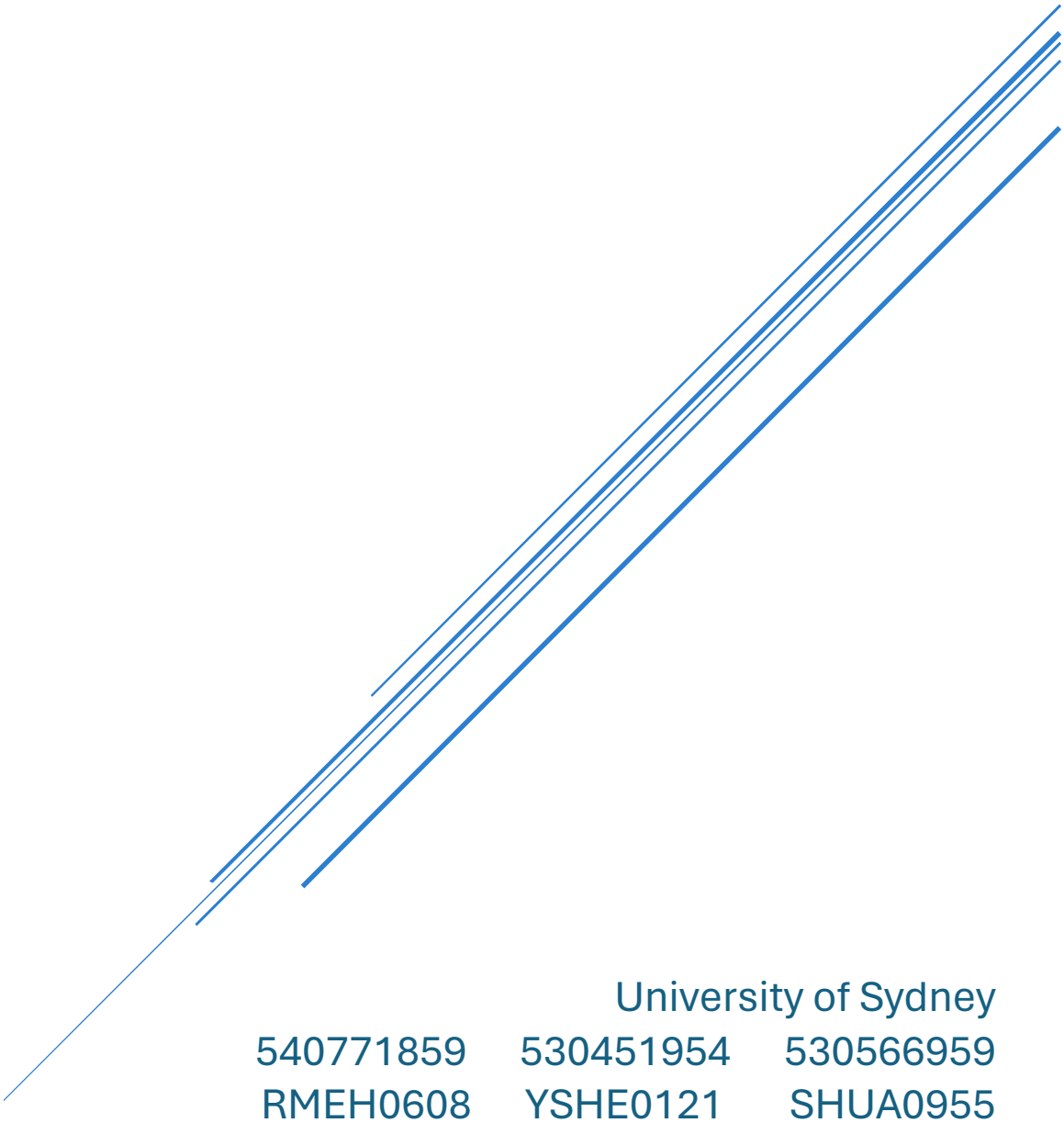
PRINCIPALS OF DATA SCIENCE

Assignment 2



University of Sydney

540771859 RMEH0608

530451954 YSHE0121

530566959 SHUA0955

1 INTRODUCTION

1.1 SET UP-RESEARCH QUESTION

Amidst the ongoing increase in price due to shortages in manufacturing parts, the second-hand vehicle market has become exponentially more important to the availability of personal transport around the world. Though, with an increased importance in this industry it is important for both buyers and sellers to understand how different factors such as the condition of the vehicle and how different models of vehicle can affect the market price and the actual sale price. Through answering which factors will determine the selling price of vehicles, sellers will be able to accurately predict the optimal value of their vehicle and buyers will be able to determine the actual price and determine if the asking price is close enough to the market value of the vehicle.

1.2 SET UP-DATASET

The dataset contains 511,661 records, distributed in 14 fields, covering a variety of data types from character brand, model, color to integer odometer, sales price and other data types. Descriptive statistical analysis of a data's yielded key finding at terms of transmission systems, the vast majority of vehicles are equipped with automatic transmissions. The data is highly consistent. Vehicle condition scores range from 1 to 49, with an average = 30.6. The data set contains vehicles ranging from almost new to poor condition.

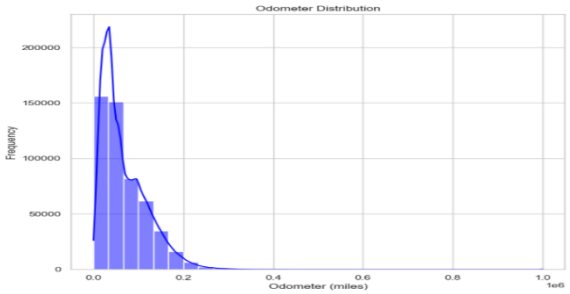


Figure 1 Odometer Distribution

The mileage distribution is mainly concentrated in low-mileage areas, with a few high-mileage vehicles and extreme outliers. Therefore, most vehicles have low mileage, and it is necessary to pay attention to the impact of a few high mileage and extreme values.

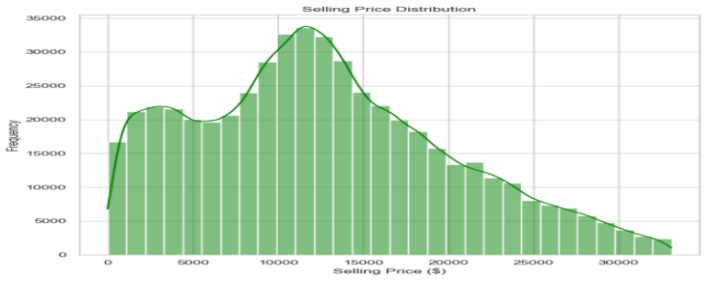


Figure 2 Selling Price Distribution

Selling prices were approximately normally distributed skewed right, indicating most vehicle prices in a mid-to-low range, but there are a small number of high-priced vehicles that increase the average price. Therefore, vehicle sales prices are mainly concentrated in the mid- to low-end, and the market also has demand for high-priced vehicles. The long tail shows that most vehicles are priced lower, with a few high-end new cars selling for well above average prices.

Faced with this data, several key challenges must be dealt with. Outliers, extremely high mileage, need to be eliminated or specially treated to avoid distorting the predictive power of the model. Secondly, the current data set shows that each column has complete values, but in practical applications it is often necessary to deal with missing data, and there must be a systematic method to estimate missing values. Furthermore, datas imbalance, especially if some brands are much more numerous than others, will weaken the model's ability to generalize to the entire market.

Geographic bias were also a factor that must be considered. If some states or regions have more vehicle data than others, the model cannot accurately reflect nationwide vehicle price trends. So, when building predictive models, it was important to ensure that these geographical biases are appropriately handled to enhance the comprehensiveness and accuracy of the model.

1.3 SETUP-MODELLING AGREEMENT

Our group will use multilinear regressions, random forest regression and decision trees to predict the actual selling prices of used car. When measuring the attributes and criteria of success, Root Mean Square Error (RMSE) and Coefficient of Determination (R ^ 2) are commonly used in regression models.

Firstly, RMSE is an indicator used to measure the difference between model predictions and actual observations. It is Mean Squared Error, the square root of MSE. The smaller the value of RMSE, the smaller the difference between the predicted results of the model and the actual observed values, and the better the fitting effect of the model. Compared to MSE, RMSE is easier to interpret because it has the same units as the original observation values.

Secondly, R ^ 2 is another commonly used measure of success, which measures whether the degree of variation of the dependent variable (target variable) can be explained by the independent variable (predictor variable). Specifically, R ^ 2 represents the proportion of explanatory variables predicted by the model to the total variation. The value range of R ^ 2 is between 0 and 1, and the closer it is to 1, the better the model fits the observed data. When R ^ 2 approaches 1, it can be confirmed that the model has a high degree of interpretation of the data, thereby proving the accuracy and applicability of the model.

Overall, RMSE and R ^ 2 provide a clear and accurate evaluation of model performance and fit level.

2 540771859 – MULTIPLE LINEAR REGRESSION

2.1 MODEL SUMMARY

To analyse and make future predictions within the second-hand market multiple-linear regression (MLR) can be used to determine how different factors influence the sale price. In simple linear regression, a line a best fit is used to determine the relationship between an independent and dependent variable (variable being predicted). The model determines the line of best-fit by minimizing the sum of the square terms (residuals) between each datapoint and the line until it reaches the smallest possible value. Then to determine the model fit, the coefficient of determination (R^2) is then calculated by the equation seen in the equation below:

$$R^2 = 1 - \frac{Error\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad [1]$$

Similarly, to a standard linear regression, MLR uses the same concepts discussed above with added dimensionality through using multiple independent features to predict one dependent variable. This causes a plane to be created to calculate the best relationship between the independent variables and the predictor variable as seen in Figure 3. Through this, the model creates coefficients on the amount of influence the independent variables affect the dependent variable and whether they are significant to the model, giving a model equation as seen below where 'beta' represents variable coefficients, 'x' represents independent variables and 'epsilon' representing the residual error.

$$MLR\ Equation: Y = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_q\chi_q + \varepsilon$$

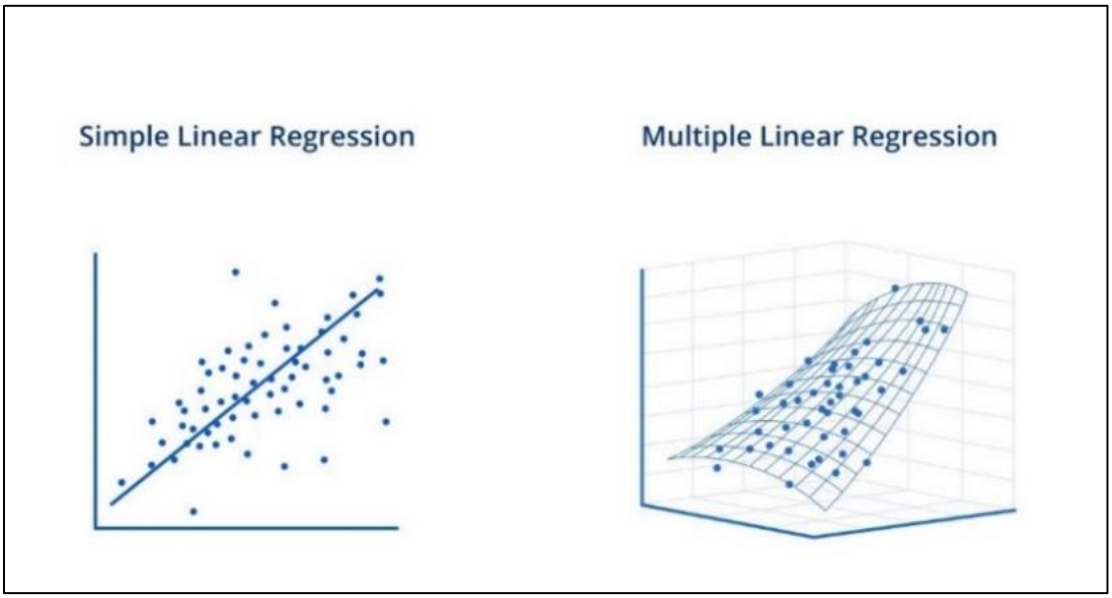


Figure 3. Simple Linear vs Multiple Linear Regression [2]

Using MLR requires certain assumptions made to the dataset, for the model to yield precise and accurate results. The first is that there exists a linear relationship between all discrete independent variables and the dependent variables. It also assumes that the residuals are normally distributed and there is no multicollinearity which means that the independent variables cannot predict each other.

Using MLR comes with certain advantages over other regression models. First, the model can easily handle multiple independent factors while also outputting the effect they have on the dependent variable. The model is also quite simple to learn and understand making it easier to hyper tune and modify for users. Though, as more factors are added, issues with amount of data needing to be processed can exponentially grow causing issues processing speeds while also increasing the risk of multicollinearity within independent variables. Also, as the model is designed to look at linear relationships, MLR does not handle categorical variables very well, resulting in them needing to be transformed through one-hot encoding or dropped which can increase the complexity of the model [3].

2.2 MODEL ALGORITHM

Due to the categorical variables present in the dataset, it was necessary to pick a method of MLR that would be easily interpretable and modifiable. For this reason, an Ordinary Least Squares (OLS) model using the 'statsmodel' library was implemented to find the best line of fit under multiple dimensions.

1. Load and Preprocess Data:
   1.1. Load data from an Excel file into a DataFrame.
   1.2. Filter rows in independent discrete variables based on distribution (e.g., 'Age', 'condition', 'odometer').
   1.3. Compute descriptive statistics for numerical and categorical variables.
   1.4. Visualize correlations and distributions of numerical variables.
2. Feature Engineering:
   2.1. Define features (X) and the target variable (salesprice) for the regression model.
   2.2. Perform one-hot encoding on categorical columns.
   2.3. Standardize numerical columns using StandardScaler.
3. Split Data into Training and Testing Sets:
   3.1. Split the dataset into training and testing sets using train_test_split.
4. Build and Fit the OLS Regression Model:
   4.1. Add a constant term for the intercept.
   4.2. Fit an Ordinary Least Squares (OLS) regression model using statsmodels.OLS.
   4.3. Print the statistical summary of the fitted OLS model.
5. Model Evaluation and Visualization:
   5.1. Calculate predictions and residuals from the fitted OLS model.
   5.2. Visualize the relationship between predicted values and residuals.
   5.3. Plot actual vs. predicted values to assess model performance.
   5.4. Compute and print model performance metrics (MSE, RMSE, R-squared).
6. Export Results to Excel:
   6.1. Save descriptive statistics and OLS results summary to an Excel file.
   6.2. Write descriptive statistics to a new sheet in the Excel file.
   6.3. Write the OLS results summary to another sheet in the Excel file.
7. Error Handling:
   7.1. Implement exception handling to manage file loading and other potential errors.

2.3 MODEL DEVELOPMENT AND EVALUATION

To analyse the second-hand vehicle market dataset, the first step is to import the data and check the data to make sure the assumptions of MLR had been met. Looking at discrete features, 'Odometer', 'Age' and 'MMR' are skewed with the distribution of condition being spread. Therefore, basic filtering was applied to remove outliers and normalise the distributions. Though it is still worth noting that after filtering, there appears to still be a slight skew amongst the discrete variables, which should be considered in the final analysis of this model. Distribution change can be seen in Appendix 8.1. Looking at the categorical variables as well, it was important to note that the make and the model of the car could essentially be seen as the same feature as each brand makes only makes their specific car models. For this reason, the make was dropped as the models would provide more useful predictions and detail. To further reduce the complexity of the model, the body, trim and seller were also removed the number of unique values for each column exceeded 50 with no normal distributions.

Looking at the output of the model, there are several interesting findings from the model. Firstly, looking at the model fit and validity, the Durbin-Watson Test which is used to determine multicollinearity in the model outputs a value of 1.9 indicating the independent variables have no autocorrelation with each other [4]. The OLS adjusted R2 value also gives a value of 0.959 indicating an almost perfect fit of the data. This is further shown by Figure 4 where the predicted and actual values have a positive linear relationship.

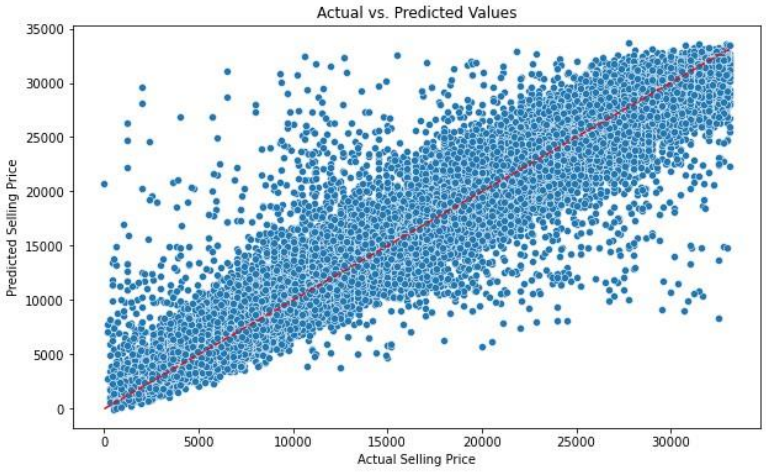


Figure 4. Predicted vs Actual Selling Price

2.4 MODEL OPTIMIZATION

Upon closer examination of the model's features, it was observed that the categorical variables exhibited p-values greater than 0.05, indicating they had an insignificant impact on the sale price. Notably, all these variables demonstrated similar p-values (0.633), suggesting they could be subjected to dimension reduction. To optimize the model, a dimension reduction strategy was implemented, starting with the removal of the 'body', followed by 'state', 'interior', and 'color'. 'Model' became significant after dimension reduction. The model's fit remained unaffected and because of these optimisations, the execution time was significantly reduced, demonstrating enhanced efficiency without compromising accuracy. The retention of the 'model' variable underscores its critical role in predicting sale price within the model.

3 530451954 - RANDOM FOREST

3.1 MODEL SUMMARY

For the vehicle_sales dataset, we're focused on predicting vehicle prices in the second-hand market. Here, we designate the continuous value of 'selling price' as the target label. To tackle this task, we employ a Random Forest regression model (Breiman, 2001). Random Forest, named so due to its composition of hundreds or even thousands of decision trees (Breiman et al., 1984), follows a distinctive approach. It only considers a subset of features in the construction process of each tree, instead of all features. Moreover, it employs bootstrap sampling to obtain bootstrapped samples with replacement. These two factors aim to ensure that each tree in the random forest is distinct, thereby enhancing the overall model performance and stability through the ensemble of diverse models.In Random Forest regression, the final output is the average of outputs from all decision trees.

This method, as an extension of the CART technique, demonstrates superior predictive performance. Random Forest regression adopts an ensemble method, specifically the bagging algorithm, to combine all generated decision trees. Based on bootstrap sampling, the bagging algorithm selects several bootstrap samples and employs the same learning algorithm. Subsequently, it aggregates the predictions of all generated base classifiers to reduce model variance and improve generalization capability. The framework for prediction using Random Forest regression is illustrated in Figure 5 (Li et al., 2018).

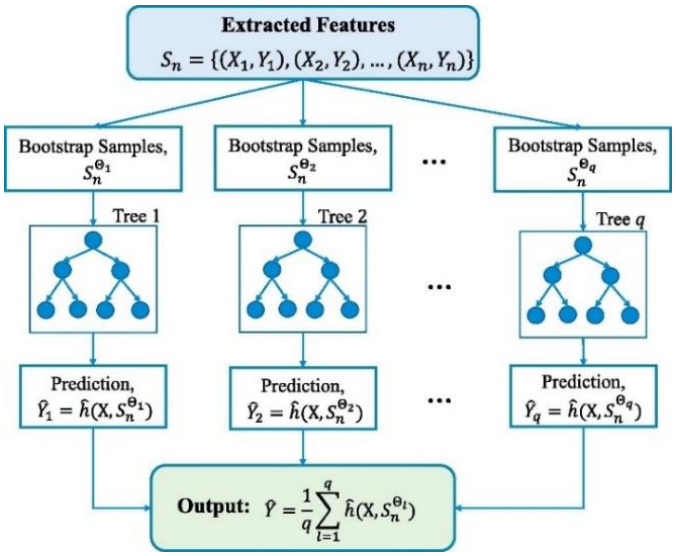


Figure 5. Random Forest Regression Construction

Bagging can effectively reduce model variance by amalgamating the prediction outcomes of multiple models, making it particularly advantageous for high variance models. Through the utilization of distinct subsets during the training phase, Bagging adeptly curtails overfitting tendencies of the model. Additionally, employing ensemble methods enhances the stability of Random Forest, shielding against disruptions that might otherwise impact prediction accuracy.

3.2 MODEL ALGORITHM AND DEVELOPMENT

After loading and simply checking the data, it is evident that there is categorical data in the dataset. For regression models, non numerical data needs to be converted to numerical types, so unique hot encoding will be applied to these columns. However, due to the potential increase in data dimensionality caused by

using unique encoding, the category data column 'seller' with excessively high unique values and no regression value was first removed. After analysis, it was found that 'model', 'trim', and 'body' were highly correlated with 'make', so they were also removed. The proportion of imbalanced categories in the dataset is close to 1, so there is no need to handle it. Using Z-core standardization can eliminate scale differences between different features, accelerate convergence, and improve model performance. The dataset is divided into training and testing sets, and during the grid search process, the training set is divided into training subsets and validation sets.

1. Loading and preprocessing data:
   -Load data from an Excel file into DataFrame.
   -Move the label column 'sellingprice' to the last position.
   -Delete columns with unique values that are too high (for example, 'seller').
   -Calculate the proportion of category imbalance
2. Feature engineering:
   -Perform one-hot encoding on categorical variables.
   -Normalize data using Z-score standardization.
3. Model building:
   -Split the data into training and testing sets:
   -Use train_test_split to split the dataset into training and testing sets.
   -Build and fit a random forest regression model:
   -Train the model on the training set.

Predict the target variables for the test set. 4. Grid search
   -Use grid search for hyperparameter tuning.
   -Train the model using the best parameters obtained through grid search.

5. Model evaluation and visualization:
   -Calculate the mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) between predicted and actual values.
   - Visualize prediction results.

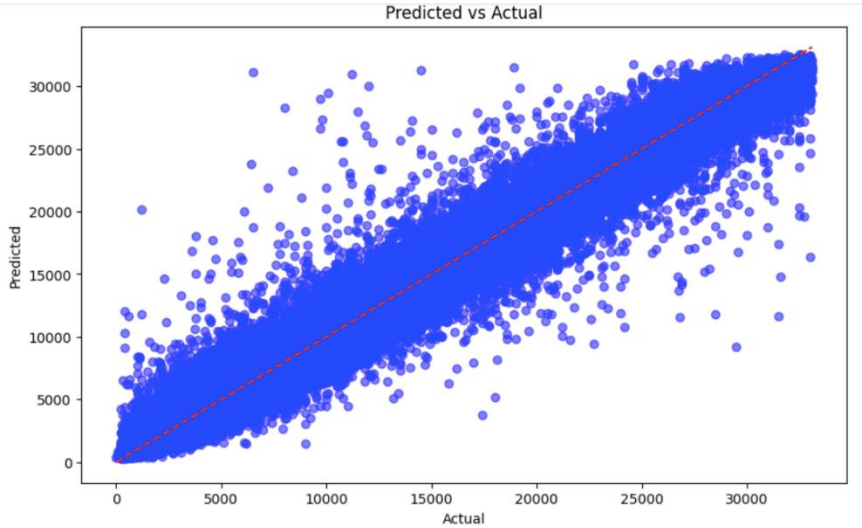


Figure 6. Scatter of Predicted vs Actual

3.3 MODEL EVALUATION AND OPTIMIZATION

3.3.1 Evaluation

For assessing the predictive performance of the regression model, the prediction results are evaluated by comparing them with the ground truth. The following are the metrics used in this task to gauge prediction quality.

1. Mean Absolute Error (MAE) MAE calculates the average absolute difference between the predicted and actual values. A lower MAE indicates more accurate predictions.

$$MAE=n1\ \sum i=1n\ |\ yi-y\hat{}\ i\ |$$

n represents the number of samples, yi denotes the actual value, and $\hat{y}i$represents the predicted value.

2. Mean Squared Error (MSE) MSE measures the average squared difference between the model's predicted values and the actual values. It squares the prediction errors for all samples, then sums them up and takes the average. MSE is a non-negative value, with smaller values indicating better predictive performance.

$$MSE=n1\ \sum i=1n\ (yi-y\hat{}\ i)2$$

n represents the number of samples, yi denotes the actual value, and $\hat{y}i$represents the predicted value.

3. Root Mean Squared Error (RMSE) Root Mean Squared Error (RMSE) is similar to MAE but penalizes larger absolute values more by giving them more weight. The larger the difference between MAE and RMSE, the greater the variance in individual errors.

$$RMSE=n1\sum i=1n(yi-y\hat{}\ i)2$$

n represents the number of samples, yi denotes the actual value, and $\hat{y}i$represents the predicted value.

The results of prediction quality:

MAE	MSE	RMSE
881.3631575835753	1732084.3336044236	1316.0867500299605

3.3.2 Grid search

Random Forest regression typically requires tuning only two parameters: the number of decision trees (n_estimators) and the maximum depth of each tree (max_depth). As the number of trees increases, the model's accuracy tends to improve, but excessively large numbers of trees can burden computational resources. Increasing the maximum depth of the trees enhances individual tree performance, but it also increases correlation among the trees.

However, while grid search is a commonly used parameter tuning method, it does not guarantee improved model performance. In some cases, grid search may be constrained by the parameter range, leading to the discovery of only local optimal solutions rather than achieving the global optimal solution.

4 530566959 - DECISION TREE

4.1 PREDICTIVE MODEL: DECISION TREE MODEL DESCRIPTION

Decision trees is a non-parametric supervised learning method used for classification and regression tasks. This model learns simple decision rules inferred from the data features. The core principle of decision trees involves segmenting the dataset into subsets through repeated divisions, targeting the most critical attributes for creating uniform segments. This method, known as optimal partitioning, involves key parameters like max_depth to control the tree's expansion and prevent overfitting, and min_samples_split, which sets the minimum number of data points required to split a node.

4.2 MODEL ALGORITHM - DECISION TREE METHODOLOGY

The decision tree methodology involves identifying the optimal split for data division, dividing the dataset at the chosen points, and recursively applying the same procedure to each resulting subset. The process terminates when reaching the tree's maximum depth or when no further division is feasible. a approach advantageous because of its simplicity the clear, easy-to-understand explanations it provides through tree visualization.

4.3 MODEL DEVELOPMENT - CONSTRUCTION PROCESS

In the decision tree construction process, data preparation does not necessarily require feature scaling, but dimensionality reduction like PCA might be employed to enhance training efficiency. Data is typically split and 70% allocated for training and the remaining 30% divided equally between validation and testing. Model configuration and parameter tuning are carried out using tools like the DecisionTreeRegressor from sklearn.tree, with adjustments to parameters like max_depth and min_samples_split based on cross-validation outcomes.

4.4 MODEL EVALUATION AND MODEL OPTIMIZATION STATISTICAL INDICATORS

Selecting key statistical indicators is critical for evaluating model performance. Metrics like MSE, RMSE, MAE, coefficient of determination R^2 provide a comprehensive view of the model's accuracy and fit.

MSE: 2402611.544849063. A higher MSE value indicates that there is some significant error in the prediction. **RMSE**: 1550.0359817917333. A larger RMSE implies that the prediction error is significant in practical applications. **MAE**: 1076.0212056029436. A lower MAE indicates that the model has higher prediction accuracy in most cases. **R^2**: 0.9565348966228917. A high value indicates that the model fits the data very well.

Analysis of Prediction Errors

The analysis of prediction errors is vital for understanding the model's performance in practical scenarios. The distribution of these errors, ideally centered around zero, should show minor and consistent deviations, indicating accurate predictions. However, significant discrepancies, such as extreme residuals, particularly at higher predicted values, suggest areas for model improvement.

Residual Analysis

Residuals' plot provides insights into the error distribution and helps identify any systematic bias or the presence of outliers. An absence of discernible trends or patterns in the residuals typically indicates no

systematic bias, while noticeable extreme residuals can highlight potential areas for model adjustment to better manage significant errors and outliers.

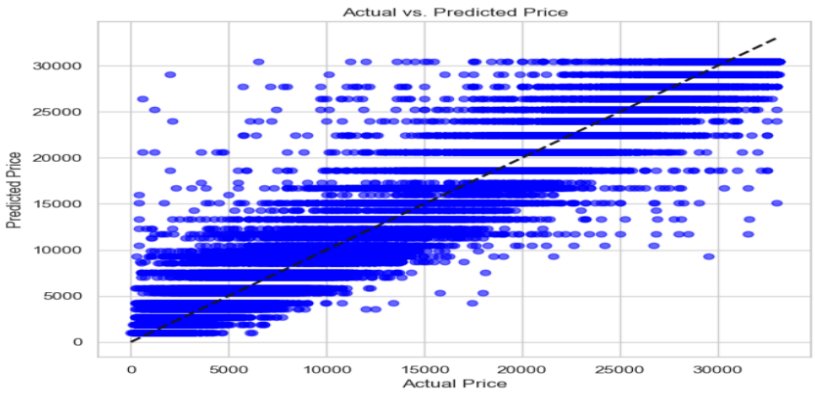


Figure 7. Actual vs Predicted Price

4.4.1 Comparison of Actual vs. Predicted Prices:

The graph depicts that the majority of the data points are clustered around the black dashed diagonal line, symbolizing a one-to-one correspondence, which illustrates that the predicted prices align very closely with the actual transaction prices. This alignment demonstrates the decision tree model's high accuracy in predictions across this particular dataset.

The patterns of predicted prices reveal that the model performs effectively across various price brackets, particularly excelling in the lower to mid-price segments where actual values tend to aggregate more tightly.

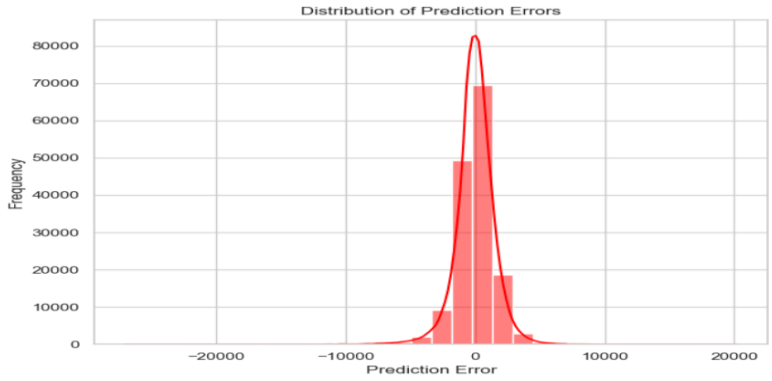


Figure 8. Distribution of Prediction Errors

4.4.2 Analysis of Prediction Errors:

The distribution of forecast errors, as shown in the chart, primarily hovers near zero and assumes a roughly normal shape, suggesting that the model's prediction errors are minor and consistent.

The errors exhibit low scatter and a peaked distribution, reinforcing that the model generally predicts accurately, with significant discrepancies being rare exceptions.

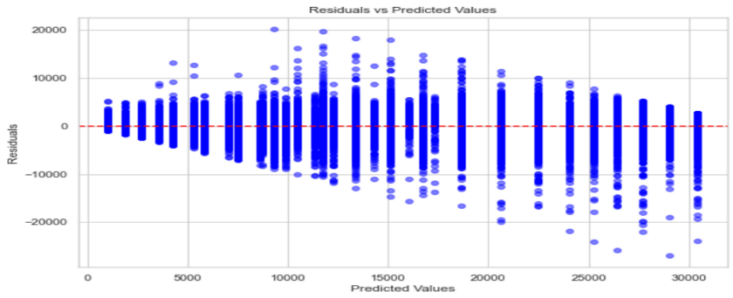


Figure 9. Residuals vs Predicted Values

4.4.3 Plot of Residuals and Predicted Values:

Error Distribution: Residuals center around the zero line, showing many accurate predictions, yet some show a broader spread, indicating larger errors. Trend Analysis: Absence of discernible trends or patterns in the residuals suggests no systematic bias in the model. Outliers: Noticeable extreme residuals occur, particularly at higher predicted values. Conclusion: The model is generally effective but could benefit from adjustments to better manage significant errors and outliers in high-value predictions.

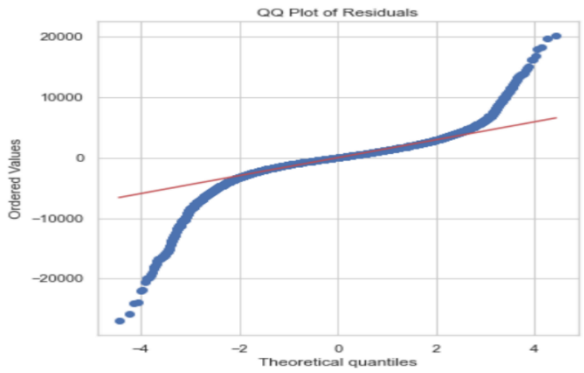


Figure 10. QQ Plot of Residuals

4.4.4 QQ Plot Analysis:

Distribution Characteristics: Data points should ideally align closely with the red reference line. However, deviations are observed at the extremes, with better alignment in the middle. Tail Behaviour: The plot reveals heavier tails, notably in the positive direction, indicating some unexpectedly large values. Conclusion: The residuals do not perfectly fit a normal distribution, showing deviations particularly at the extremes. This suggests the need for more robust handling of extreme values or revaluation of model assumptions.

4.3 Conclusion

The comprehensive analysis of used car sales data using the decision tree model demonstrates the robustness of the model in predicting used car prices across various conditions and market variables. By effectively managing outliers and exploiting inherent patterns within the dataset, the model provides reliable price predictions, enhancing market transparency and efficiency in the used car industry.

5 DISCUSSION AND CONCLUSION

The three models Multiple Linear Regression (MLR), Decision Trees (DT) and Random Forest Regression (RFR) have all been excellent choices when fitting the second-hand vehicle market data which could be used for analysis and sale price predictions. Though, with any machine learning model, there are certain benefits and disadvantages to each model which can determine their accuracy and fit of the model.

Starting with MLR, the model itself suffered from a few issues when fitted with the data. Firstly, starting with the large number of categorical variables caused the model to have a very large data input which meant that all the categorical variables were no longer significant to the model and caused a complex dimensionality increasing the processing speed of the dataset. Having to apply dimension reduction to all except 'model' meant that a lot of the data was removed before the model became feasible for use. Another factor as well was that the original assumptions for the dataset meant that the independent variables should be normally distributed. MMR could be argued to have a close to normal distribution, Age, Condition and Odometer all had slight skews which could influence the accuracy.

Though MLR suffered from these issues, it did have its advantages. Firstly, the model was easily able to determine the affect each factor has on the sale price and if the feature was significant to the prediction while also being able to easily identify outliers and extremities. The model also had a Durbin-Watson value of close to 2, indicating no multicollinearity in the model giving an R^2 value >0.9. Though, this large R^2 value could also indicate an issue with overfitting and requires more hyper-parameter testing.

Random Forest

Advantage:
High performance: Random forests can achieve good performance in many situations, whether in classification or regression tasks.

Robustness: Random forests have good robustness against noise and missing values in data, can handle many input features, and are not easily overfitting.

Interpretability: Compared to a single decision tree, random forests are easier to understand and interpret because they are an integrated model composed of multiple decision trees.

Limitation:

Memory consumption: Random forests require the construction of multiple decision trees, which may consume a significant amount of memory when dealing with large datasets.

Long training time: Due to the need to construct multiple decision trees, the training time of a random forest may be longer than that of a single decision tree.

Slow prediction speed: Although the prediction speed of random forests is usually fast, the prediction speed may be slower compared to some simple models (such as linear models).

Decision Tree

Advantage: It is easy to understand explanations, tree visualization, and capable of processing numerical category data. There is no need for big data preprocessing and normalization.

Limitation: A decision tree model is prone to overfitting. The tree is too deep, and the data is too complex. It cannot work well for certain types of nonlinear problems. Vulnerable to small changes in data, resulting in poor model stability.

Table 1. Comparison of Model Fits

Model	RMSE	R^2
Multiple Linear Regression	**1312.20**	**0.9579**
Random Forest	**1316.09**	**0.9686**
Decision Tree	**1550.04**	**0.9565**

Exploring the three models, the 'Random Forest Regressor' appears to be the best model for the prediction of selling price of vehicles. Firstly, when compared to MLR and DT, RFR does a better job at keeping the model simple as the handling of overfitting and categorical variables is easier. Though the processing time for RFR is long as it iterates over many different decisions trees, the model is more robust and can easily pick up on trends on the data. This in theory matched with the similar R^2 and RMSE values as shown in Table 1. Indicate that it is the optimal model for analysing and determining vehicle selling price.

6 REFERENCES

[1] I. Valchanov, "Sum of Squares: SST, SSR, SSE," 365 Data Science, 3 July 2023. [Online]. Available: https://365datascience.com/tutorials/statistics-tutorials/sum-squares/. [Accessed 6 May 2024].

[2] D. Sonar, "Understanding Linear Regression: The Basics," LinkedIn, 23 October 2023. [Online]. Available: https://www.linkedin.com/pulse/understanding-linear-regression-basics-divyesh-sonar-snv4c/. [Accessed 6 May 2024].

[3] E. C. Fein, J. Gilmour, T. Machin and a. L. Hendry, "Section 5.3: Multiple Regression Explanation, Assumptions, Interpretation, and Write Up," in *Statistics for Research Students*, Minnesota, University of Minnesota, 2022, pp. 35-47.

[4] W. Kenton, "Durbin Watson Test: What It Is in Statistics, With Examples," Investopedia, 27 May 2023. [Online]. Available: https://www.investopedia.com/terms/d/durbin-watson-statistic.asp#:~:text=What%20Is%20the%20Durbin%20Watson,autocorrelation%20detected%20in%20the%20sample.. [Accessed 6 May 2024].

[5] J. F. R. O. C. J. S. Leo Breiman, Classification and Regression Trees (1st ed.), New York: Taylor & Francis Group, 1984.

[6] L. Breiman, "Random Forests," *Machine Learning,* pp. 5-32, 2001.

[7] L. B. Breiman, "Bagging predictors," *Machine Learning,* vol. 1, pp. 123-140, 1996.

[8] Y. Z. C. B. M. N.-M. E. C. J. C.-W. v. d. B. P. V. M. J. &. O. N. Li, "Random forest regression for online capacity estimation of lithium-ion batteries,," *Applied Energy,* vol. 232, no. 0306-2619, pp. 197-210, 2018.

7 APPENDIX

7.1 540771859 – BEFORE MODEL OPTIMIZATION

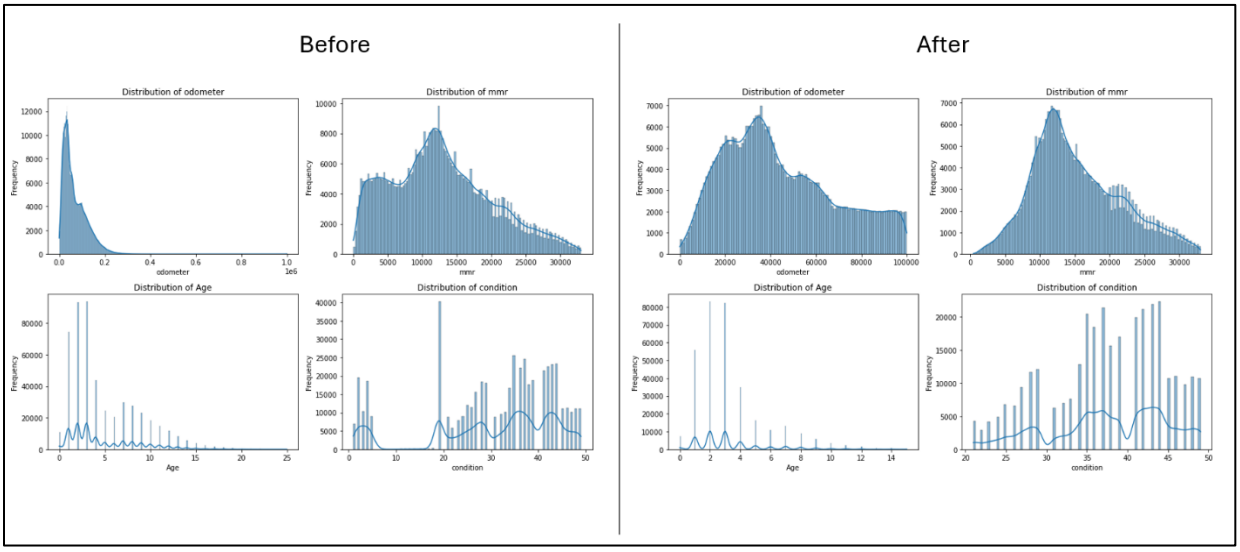


Figure of Before and After Data Transformation

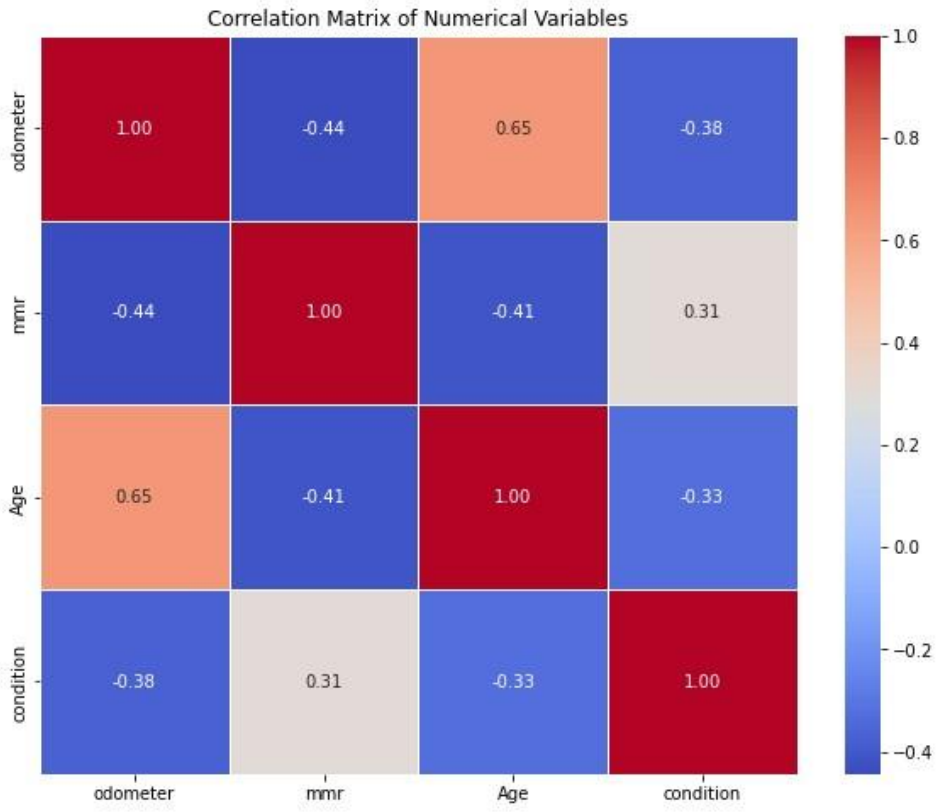


Figure of Correlation Matrix of Numerical Values

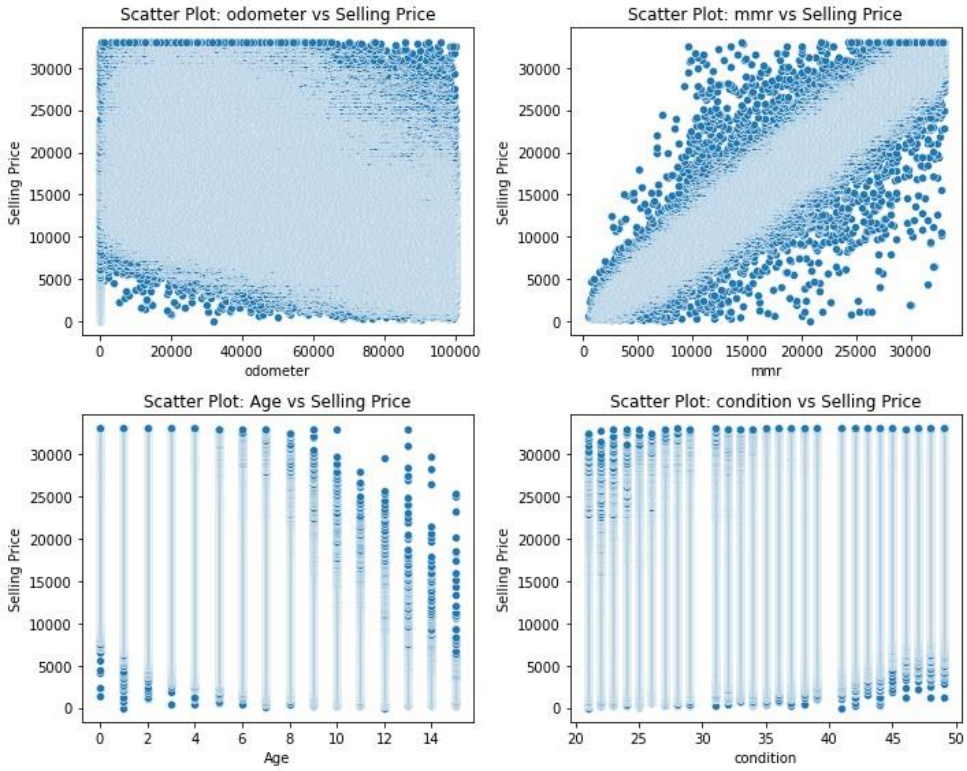


Figure of Discrete Features vs Selling Price

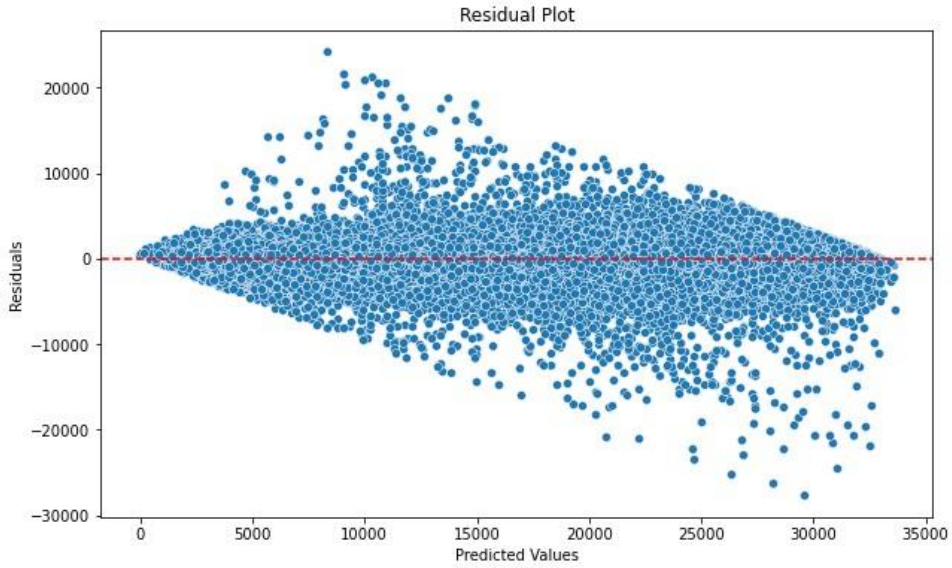


Figure of Residual Plot