Optimizing Used Car Prices Using Machine Learning

Andrew Perez

Western Governors University

Table of Contents

A Project Overview

A.1 Research Question

Is it possible to accurately predict the price of a used vehicle with machine learning by utilizing a random forest model? If this is possible, it could offer dealerships optimized pricing which could increase sales and potentially raise profit margins.

A.2 Scope

In Scope:

Data Collection & Cleaning: Collection and cleaning of data, addressing missing values and outliers.

Model Development: Building a predictive model using the random forest algorithm, with feature selection and engineering for optimization.

Model Testing & Validation: Division of data for testing, testing against unseen data, and statistical analysis for model validation.

Statistical Analysis: Comprehensive statistical evaluation of model predictions using key metrics.

Hyperparameter Tuning: Iterative optimization of model hyperparameters for enhanced predictive accuracy.

Documentation: Documentation of analysis process, cleaning steps, and modeling techniques.

Out of Scope:

Cross Model Comparison: Comparative analysis involving multiple machine learning models (linear regression, support vector machines, gradient boosting, etc.)

Model Updating: Implementation of ongoing data collection, model training, and tuning.

External Data Integration: Incorporating additional external data sources beyond the initially identified dataset.

A.3 Solution

Using the CRISP-DM methodology (B.2), this project developed a regression model utilizing the random forest algorithm. The model underwent training and testing using a clean dataset of used vehicle prices. According to Car Dealer Magazine, the average profit margin for a used car sale is between 12 and 15% (Baggott, 2021). To have a viable solution, the model needed a MAPE value less than that average margin, so success was determined by comparing the model's MAPE value to a goal of 5% or less. MAPE, representing the percentage difference between the model's price predictions and actual values, serves as a crucial metric for accuracy and precision.

Tools Used:

Python: An open-source, versatile programming language with a vast set of libraries. The following libraries were used in this project.

pandas: Data cleaning, preprocessing, and analysis

NumPy: Scientific computing

SciPy: Used in combination with NumPy. In this case, for outlier identification through z-scoring.

sci-kit-learn: Machine learning algorithms, metrics, and model selection.

Matplotlib: Basic visualizations

Jupyter Notebooks (IDE): Interactive, web-based IDE that allows for the combination of code, analysis, and documentation in one environment.

B Project Execution

B.1 Project Plan

The following plan was proposed for this project:

Goal: Build a random forest model that can accurately predict a used vehicle's price while maintaining a mean absolute percentage error (MAPE) of 5% or less.

Objective 1: Collect, clean, and encode a used vehicle dataset to train and test the model.

Deliverable 1: A clean, complete, and encoded pandas data frame.

O2: Develop an optimized random forest model through hyperparameter tuning and feature importance analysis.

D2: A random forest model that has been exhaustively tuned and optimized.

O3: Communicate results in a concise and interpretable report.

D3: A written and visual report that summarizes project results.

The plan's objectives were specific and concise enough so were met without variance. Unfortunately, the project's goal was not met as the model's performance did not meet our success criteria.

B.2 CRISP-DM Methodology Application

1. **Business Understanding (Analysis Phase)**

   Established a comprehensive understanding of the business problem, a potential solution, and criteria to determine the solution's viability.

2. **Data Understanding (Analysis Phase)**

   Established an understanding of the dataset, quality issues, and transformation needs.

3. **Data Preparation (Design Phase)**

   Prepared a clean dataset to be used in the modeling phase.

4. **Modeling (Design & Development Phase)**

   Developed, trained, and tuned a random forest model.

5. **Evaluation (Testing Phase)**

   Assessed model performance against success criteria defined above.

6. **Deployment (Implementation Phase)**

   The model was unsuccessful in meeting or exceeding our MAPE goal, therefore deployment was not initiated.

B.3 Execution Timeline

The following timeline summarizes the execution effort which took 19 days in total:

Milestone	**Duration (days)**	**Projected Start Date**	**Anticipated End Date**	**Actual Start Date**	**Actual End Date**
Collection/Understanding	1	12/11/2023	12/12/2023	12/11/2023	12/11/2023
Cleaning/Preparation	2	12/13/2023	12/14/2023	12/12/2023	12/13/2023
Model Creation/Training	1	12/15/2023	12/15/2023	12/14/2023	12/14/2023
Model Testing/Evaluation	1	12/16/2023	12/18/2023	12/15/2023	12/15/2023
Report Compilation	6	12/19/2023	12/22/2023	12/16/2023	12/21/2023
Stakeholder Communication	8	12/26/2023	12/26/2023	12/22/2023	12/29/2023
Project Closure/Completion	1	12/27/2023	12/27/2023	12/30/2023	12/30/2023

The early milestones were executed ahead of schedule due to the cleanliness of the dataset, however, the later milestones were behind schedule, largely due to holiday plans and job interviews.

C Data Collection

C.1 Collection Process

The initial dataset (I.2) proposed was used for this project as planned. The dataset was directly downloaded from Kaggle and the data itself was scrapped from Dubizzle using a Python library called Beautiful Soup. Since the dataset was publicly available, anonymous, and pre-collected, there were no significant issues with its collection or governance.

C.2 Advantages/Limitations

Advantages:

Completeness
Of the 20 features, only one is missing 10% or more of its data. Every other feature has 99%+ complete and valid data. This level of completeness allowed for a comprehensive analysis and limited any biases that could have arisen in the cleaning/preprocessing of the data.

Relevance

This dataset remains relevant as it was only collected a year ago (12/05/2022).

Size

This dataset has 9170 entries and 20 features. Those dimensions were sizable enough to train and test the random forest model.

Feature List

This dataset includes up to 19 features that were used as predictors, from transmission type to color.

Limitations:

Regional

The dataset collected from the UAE reduces its relevancy to US dealerships.

C.3 Practical Significance

To assess the practical significance of the model, a triad of the metrics listed above, MAE, MAPE, and RMSE, will be employed. In this analysis, MAE will represent the dollar amount our predictions our off by (under and overpricing.) MAPE will provide a scalable measure for that difference and will be compared to the dealership's average profit margin. If our MAPE value far exceeds the dealership's average profit margin, our solution will not be viable or successful. Lastly, RMSE is like MAE as it provides a tangible difference in pricing, but it also penalizes larger errors, which is crucial for a model that could affect the dealership's profit/loss.

C.4 Visualization Tools

With the help of Python's library Matplotlib, I will present my findings with the following visualizations:

Horizontal Bar Chart: Top Features and Their Importance on Pricing, highest to lowest

Scatter Plot: Predictions vs Actual Values

D Data Extraction/Preparation

D.1 Extraction

The dataset was downloaded directly from Kaggle in .csv format via a Chrome web browser. It was then uploaded to a Jupyter Notebook environment for cleaning, preparation, and modeling.

D.2 Preparation

Since the dataset was relatively clean, only a few steps were required to clean and prepare the data for modeling. The steps taken are summarized below, but are described in more detail in the code listed under Appendix I.1:

Cleaning: Fill in missing data points with the mean, exclude entries with missing data points that are not fillable, and remove irrelevant columns.

Formatting: Create dummy columns for nominal categorical features and encode ordinal categorical features with numerical ranges. Ensure datatypes are int or float.

Conversion: Convert price and kilometers to USD and miles.

Normalizing: Remove outliers in miles, price, and year based on z-scores greater than 3.

The following tools were used to prepare this dataset for modeling:

Python:

pandas: Cleaning, Formatting, Conversion and Normalizing

NumPy: Cleaning

SciPy: Normalizing

Jupyter Notebooks (IDE): Cleaning, Formatting, Conversion, and Normalizing

Python was ideal for this project because of its vast set of libraries that provide statistical functions and simple tools for data manipulation. Since a Jupyter Notebook is iterative, it allowed for the preparation process to be taken step by step, which minimized errors. Both tools assisted in cleaning and preparing the dataset efficiently and accurately.

Overall, the cleaning and preparation process created a complete, normalized dataset with model-ready datatypes (int/float) and values (encoded/numerical). With this new dataset, training and testing the random forest model will entail only a few lines of code.

E Data Analysis

E.1 Methods/Models

This project employed predictive analytics to forecast a used vehicle's sales price using a machine learning model. There were various statistical metrics used to assess the model, but the key metric for success was the Mean Absolute Percentage Error (MAPE).

Predictive analytics was an ideal choice because we used historical data and various features/predictors to determine an optimal price. We attempted to find a future answer by forecasting previous results. As for MAPE, it was ideal for this project because it is an interpretable metric that can be compared across various models and datasets if needed as it is scale-independent.

Below you will find the specific model used and the metrics used for assessment:

Model:

- **Type:** Supervised Regression
- **Algorithm:** Random Forest
- **Success Criteria:** MASE < .05
- **Metrics:**

- **Mean Absolute Error (MAE):** Average of the absolute difference between predicted and actual values. This metric does not consider the direction of errors (over/underpriced), only their magnitude.
- **Mean Absolute Percentage Error (MAPE):** Percentage difference between predicted and actual values. This allowed us to see the accuracy of our model regardless of the dataset and/or scale.
- **Mean Absolute Scaled Error (MASE):** Compares the performance of a forecasting model to that of a naive method. This metric considers both the direction and magnitude of errors. It pairs well with the MAPE to provide an interpretable measure of accuracy.
  - Scale of 0.0 to 1.0+, with values less than 1.0 indicating that the model outperforms a naive method.
- **Mean Squared Error (MSE):** Average of the squared differences between predicted and actual values. This metric penalizes larger errors.
- **Root Mean Squared Error (RMSE):** The square root of MSE, provides an output that is the same as the unit of the target for more interpretable measure.
- **R^2:** The proportion of the variance in the target that is predictable from the included features.
  - Scale of 0.0 to 1.0, where a higher value indicates a better fit of the model to the data (1.0 = 100% of the target's variance can be predicted by the included features)

E.2 Advantages/Limitations

Advantages:

- **Handles Complex Relationships:** Ability to predict continuous numerical values (prices) based on inputted features/predictors even if those predictors have complex, non-linear relationships.
- **Utilizes Numerical and Categorical Data:** Both numerical data, like miles and transmission, and categorical data, like color and body type, can be utilized in the model's prediction.
- **Built-In Feature Importance:** The model can provide easily interpretable estimates of each feature's importance in producing a prediction.

Limitations:

- **Computationally Complex:** Training many trees in a random forest can be computationally intensive, especially for large datasets. This limits the scalability of the model as well as its cost benefit.

E.3 Method Application

Random Forest Model

- **Steps:**
  1. Split clean dataset into training and testing sets.
  2. Fit the model with the training set.
  3. Validated model against testing set.
  4. Evaluated performance of the base model through various metrics (E.1).
  5. Tune hyperparameters to optimize performance.
  6. Repeated 1-5 until optimal performance was realized.

- **Requirements:**
    1. A Clean, Complete Dataset
    2. Encoded Categorical Features
    3. Numerical Target
- **Verification:** All requirements were completed and verified in the data preparation step (D.2) of this project.

F Results

F.1 Statistical Significance

To find the optimal model to test our null hypothesis against, the random forest model was iterated 8 times, and here are the results from each iteration:

Metrics	FD-B	FD-R	FD-G	IQR-B	IQR-R	IQR-G	FD-G-10	IQR-G-10
MAE ($)	17,272.82	14,764.14	14,764.70	6,633.11	6,358.13	6,334.14	14,645.48	6,474.08
MAPE (%)	37.24	31.83	31.83	24.46	23.44	23.36	31.58	23.87
MASE	.34	.29	.29	.56	.53	.53	.29	.54
RMSE ($)	28,785.44	25,700.32	25,707.13	8,326.56	8,165.68	8,059.27	25,627.09	8,430.14
R^2	.76	.81	.81	.36	.39	.40	.81	.35

FD = Full Dataset	**IQR** = IQR Dataset	**10** = Top 10 Feature Reduced
B = Base Model	**R** = Random Search Tuned Model	**G** = Grid Search Tuned Model

Unfortunately, none of the iterations allowed us to reject our null hypothesis which states that a random forest model cannot accurately predict a used vehicle's price while maintaining a mean absolute percentage error (MAPE) of 5% or less. The closest model was IQR-G with a value of 23.36%. It seems that to make that gain, the IQR-G model sacrificed its MASE and R^2 values (.53/.40 compared to the FD-G model's .29/.81.)

F.2 Practical Significance

From a practical standpoint, our model was also not successful. Our lowest values for MAE and RMSE were also found in the IQR-G model with values of $6,334.14 and $8,059.27. Overpricing a vehicle by $6-8k might be more profitable, but it's likely to lead to less business as customers will flock to dealerships with more reasonable prices. On the other hand, underpricing a vehicle by the same amount will cause a loss that will be hard to offset with the average 12-15% profit margin that most dealerships make on used vehicles (Baggott, 2021).

F.3 Overall Success

Overall, this project was not statistically or practically successful. We are unable to reject the null hypothesis and deploying this model for pricing would be at best ineffective and at worse very detrimental to the dealership's profits.

G Conclusion

G.1 Summary

The random forest algorithm was exhaustively trained, tuned, and examined in this project and it failed to predict used car prices with any viable accuracy. It didn't matter if the full dataset was used, if features were removed, or if outliers were excluded. Tuning the model's hyperparameters did improve accuracy, but not enough to make this model a viable solution.

Our highest-performing model, which trained on the dataset's interquartile range (IQR) and was optimized by cross-validating its hyperparameters by grid search, reached a MAPE value of 23.36%. That model also posted a .53 MASE score and a .40 R2 score. This shows that while our model outperforms the accuracy of a naïve prediction, it does not fit our data well, as seen in I.3, and comes nowhere near close enough to be deployed in real-world scenarios.

In summary, the null hypothesis cannot be rejected as we were unable to produce a random forest model that could predict a used vehicle's price while also maintaining a MAPE value of 5% or less.

G.2 Effective Visual Storytelling

The following visualizations were created in a Jupyter Notebook using Python's library Matplotlib. They effectively summarize this project's conclusion and give a visual representation of our model's performance as well as key features in used vehicle pricing.

Model Accuracy

Description: Two scatter plots showing the accuracy of the IQR-G and FD-G models. There are three colorblind-accessible colors in the plot. One to indicate when the model has overpredicted a price, one to indicate when the model has underpredicted, and one to show when the model was within 5% of the actual price. There is also a fit line to show how well our models fit the dataset it was trained on.

Purpose: This visualization is important because we can see the differences between our highest-performing model iterations, including their accuracy. Despite the IQR-G model posting a lower MAPE score, it was only able to predict a handful of prices within 5% of the actual price. The fit for the IQR-G was significantly off as well. Comparing the IQR-G plot to the FD-G plot shows the trade-offs we encountered in optimizing our model for the MAPE metric. In the FD-G plot, we see a model with a good fit, but even fewer predictions within 5% of the actual prices.

Model Feature Importance

Description: A horizontal bar chart that shows the top 20 features for predicting a used vehicle's price based on importance in our IQR-G model.

Purpose: This visualization provides us with the key predictors in used vehicle pricing. Even with an inaccurate model, we can see the top features are as one might assume, miles, horsepower, year, and cylinders. These features relate to a vehicle's performance instead of its description. Our descriptive features may have created too much noise. Perhaps with a dataset that has more performance-based features, we could improve our model.

G.3 Courses of Action

There are two courses of action the dealership can take in response to this project and its results. The first and most simple is to maintain their current pricing model and reject incorporating machine learning into how they price a vehicle. The second is to take this project's results and work on improving them through experimentation with other machine learning algorithms like linear regression, support vector machines, or gradient boosting. This second course of action could also focus on different datasets that incorporate more performance-based vehicle features instead of descriptive features like color and body type. The dataset used in this project was of vehicles listed on a public marketplace so it's possible with internal data and a more exhaustive feature list, our prediction capabilities could improve.

H Presentation

Panopto Video

I Appendices

I.1 Code

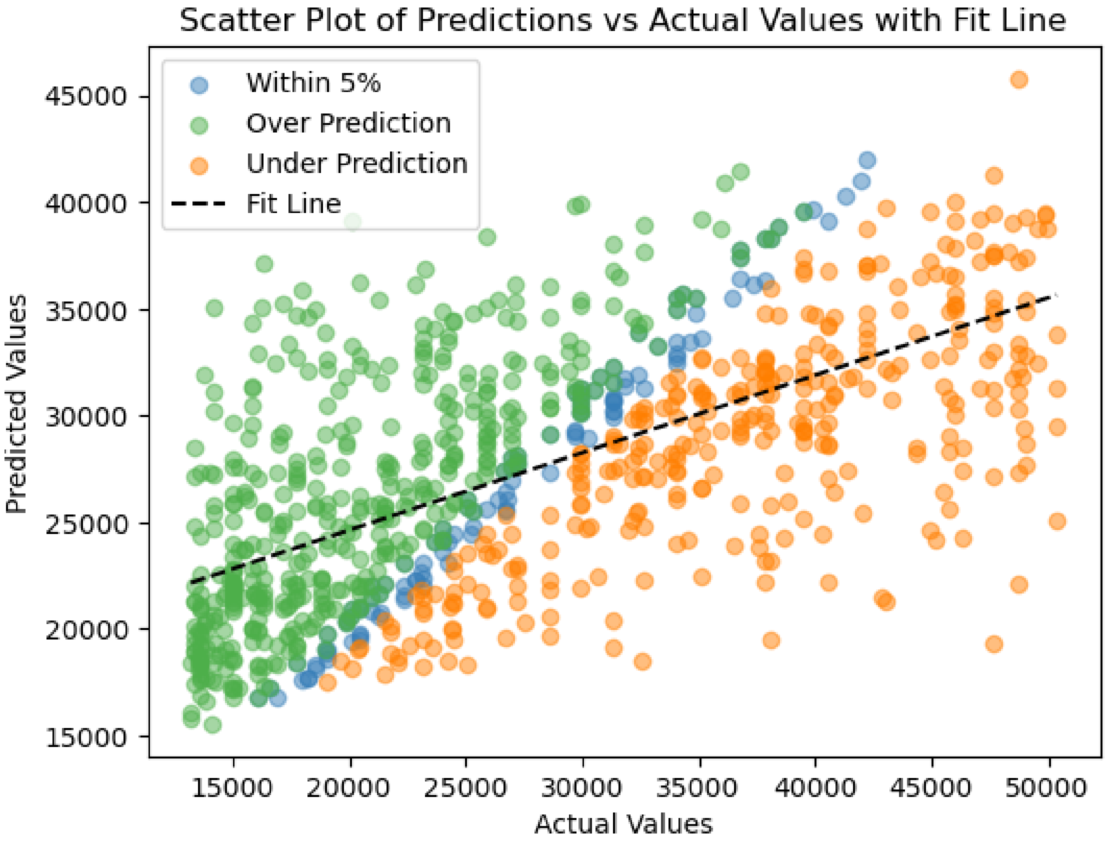
GitHub - Capstone

I.2 Dataset

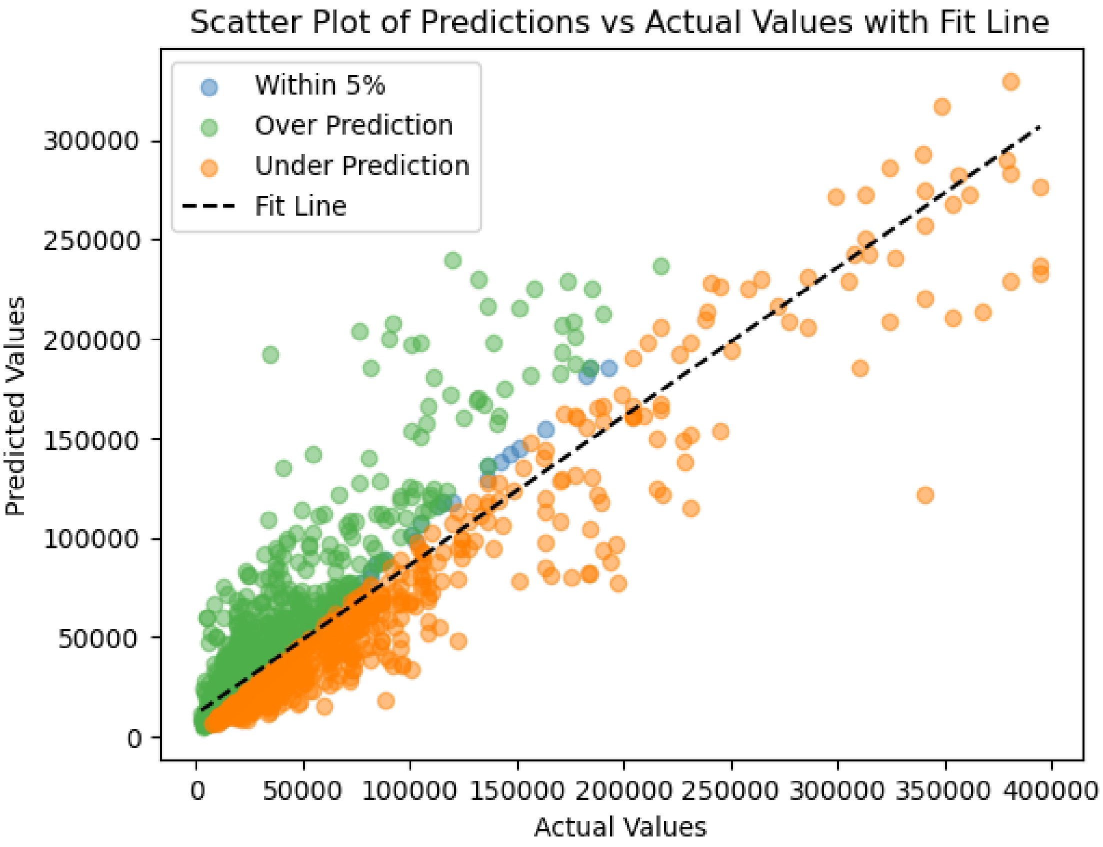
Dubizzle Used Car Sales Data (2022) – Ali Hassan

I.3 Model Accuracy Visualization

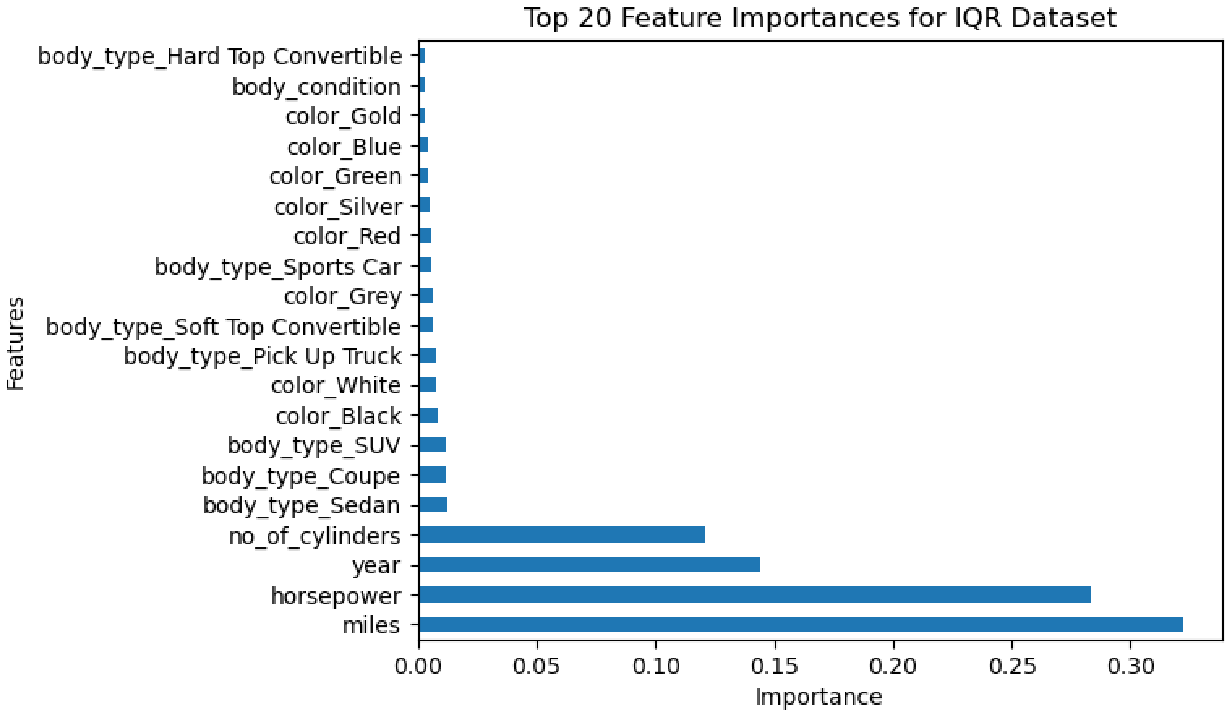
IQR-G:



FD-G:


Scatter Plot of Predictions vs Actual Values with Fit Line
Within 5%
Over Prediction
Under Prediction
Fit Line
Predicted Values
0
50000
100000
150000
200000
250000
300000
Actual Values
0
50000
100000
150000
200000
250000
300000
350000
400000

I.4 Model Feature Importance Visualization



J References

Baggott, J. (2021, July 20). *How much profit do car dealers make on new and used cars? here's the perception versus reality*. Car Dealer Magazine. https://cardealermagazine.co.uk/publish/how-much-profit-do-car-dealers-make-on-new-and-used-cars-heres-the-perception-versus-reality/225342