CIDM 6355 FINAL PROJECT

**Used Car Price Prediction Model**



**CIDM 6355 Project**

**PROFESSOR: DR. LIANG CHEN**

**Team Name:** Team E

**Team** **Number:** 5

**Project Proposal:** Used Car Price Prediction

**Team Members:**

|  |  |
| --- | --- |
| 1. Joshua Tarrant | 2. Farzaneh Noroozi |
| 3. Pradnya Bhere | 4. Joshua John |
| 5. Brian Doss | 6. Mallika Alan |
| 7. Trevor Hofmann |  |

# EXECUTIVE SUMMARY

In the ever-evolving landscape of the automotive industry, the ability to predict used car prices with precision is indispensable for maintaining competitive advantage. This report encapsulates the findings from the Used Car Price Prediction Project, which aimed to harness advanced data analysis and machine learning techniques to forecast the prices of used vehicles.

**Project Overview** The used car market, with its complex interplay of factors influencing pricing, presents a substantial opportunity for data-driven decision-making. This project, grounded in a comprehensive dataset from Kaggle, sought to model these factors and predict price ranges, offering valuable insights for buyers and sellers alike.

**Research Methodology & Findings** Our approach combined meticulous data preparation with a suite of data mining techniques, including regression analysis, decision trees, neural networks, and Naïve Bayes classification. The models were rigorously evaluated for accuracy, precision, recall, and F-measure, with Naïve Bayes and Neural Networks demonstrating superior performance.

**Implications for Stakeholders** The predictive models developed through this project serve as potent tools for various stakeholders, enabling car dealerships to optimize pricing strategies and empowering consumers with informed purchasing and selling insights.

**Challenges and Limitations** While the project yielded promising results, it also encountered limitations such as potential data quality issues and the static nature of the dataset, which may not reflect market dynamics over time.

**Future Recommendations** To enhance the robustness and applicability of the predictive models, we recommend expanding the dataset and incorporating more dynamic market and economic indicators. Advanced machine learning techniques could further refine the predictive accuracy.

**Concluding Thoughts** The Used Car Price Prediction Project lays a solid foundation for predictive analytics in the automotive sector, offering stakeholders a data-centric approach to navigating the used car market's complexities. With ongoing refinement and adaptation to market shifts, this project is poised to become an invaluable asset for the industry.

**I. Introduction**

**Industry Background**

In today's dynamic automotive market, characterized by a thriving trade in used cars, the ability to accurately forecast the prices of these vehicles has become a vital necessity. Understanding the intricate web of factors that influence used car pricing is crucial for both buyers seeking value and sellers aiming for optimal returns on their automotive assets. This project utilizes a comprehensive dataset and advanced data analysis techniques to predict the price range of used cars.

**Context:**

The used car market is a complex and multifaceted industry, with pricing at its core. Precise price predictions are critical for navigating this landscape. In response to these demands, our project aims to leverage data mining techniques effectively for predicting used car prices. This proposal outlines the problem statement, data source, and project plan for this endeavor.

**Problem Statement**

The automotive industry is marked by constant evolution, driven by shifting consumer preferences, economic factors, and environmental concerns. One of the primary challenges within this industry is accurately predicting used car prices. Buyers seek guidance to make informed decisions, while sellers aim for competitive pricing.

**Stakeholders**

Our project directly serves two key stakeholder groups:

* **Buyers**: Individuals seeking to purchase used cars seeking insights into how various car attributes impact prices.
* **Sellers**: Individuals or businesses looking to sell used cars aiming to set competitive prices to maximize their chances of sale.

**Objective & Motivation**

At the core of this project is the development of predictive models capable of estimating the price range of used cars based on a diverse set of attributes. Our intent is twofold. Firstly, to provide prospective buyers and sellers in the used car market with a treasure trove of invaluable insights. And, secondly, to uncover the attributes that wield significant influence in the intricate art of pricing used cars.

The motivation behind this project stems from the desire to empower both buyers and sellers in the used car market. By delivering accurate price predictions, we aim to facilitate well-informed decisions, enhance pricing strategies, and contribute to a more transparent and efficient marketplace.

**Opportunity/Challenge**

The central challenge of our project revolves around the development of a robust prediction model capable of accurately estimating used car prices based on a diverse set of attributes. We must navigate a wide spectrum of price points and convert this variability into a meaningful classification task.

**Data Description**

Our analysis centers on a meticulously curated dataset comprising 4,009 individual records, each offering a unique glimpse into an automobile's characteristics. These records encompass twelve key attributes, each holding critical information about these vehicles. These attributes include details such as the car's brand, model, model year, mileage, engine type, transmission, exterior condition, interior colors, accident history, and the coveted clean title status. Central to our predictive journey is the categorical target attribute, "Price range prediction," which is essential for our analysis.

**Data Overview:**

We analyzed a Kaggle-sourced dataset teeming with insights about pre-owned vehicles. Comprising 4,009 entries and featuring 12 distinct attributes, our dataset paints a comprehensive picture, encompassing variables ranging from make and model to mileage and accident history. These diverse data points converge to inform our target variable: the predicted price range, which serves as the cornerstone of our predictive model.

**Data Source**: The backbone of our project, the used car dataset, is procured from Kaggle— a reputable source of data in data science. Kaggle's commitment to data integrity and reliability underpins the trustworthiness of our analysis. For those interested in delving into the data themselves, the dataset is readily available for exploration and download via the following link: [Used Car Price Prediction Dataset](https://chat.openai.com/c/insert-link-here).

**II. Research Methodology & Modeling**

**Systematic Research Methodology and Modeling Approach:** Our research methodology constructs precise and reliable predictive models for used car price ranges. This process is underpinned by the diligent collection of our dataset.

**Data Preparation and Wrangling:** Our research journey commences with a critical phase of data preparation and wrangling, where we meticulously address the following key tasks:

1. **Data Collection:** We meticulously gathered a dataset comprising 4,009 records and 12 attributes, each providing essential insights into automobiles.
2. **Data Cleaning:** We employ rigorous data cleaning to ensure the pristine quality of our dataset. This encompasses addressing missing values, handling outliers, and rectifying data inconsistencies.
3. **Attribute Selection:** To streamline our predictive analysis, we identify and select attributes that significantly contribute to our primary task of predicting car prices, thoughtfully excluding those that do not.
4. **Descriptive Statistics:** We utilize a range of tools, including Microsoft Excel, RapidMiner, and potentially R, to uncover central tendencies and data distributions.
5. **Data Visualization:** Data visualization techniques are employed to gain deeper insights into the relationships between attributes and our target variable, "Price range prediction." Crafted graphs and plots illuminate hidden patterns and trends within our dataset.

**Data Mining Techniques:** For constructing effective predictive models for used car price prediction, we leverage a diverse set of data mining techniques, including:

* **Regression Analysis:** Traditional regression models, such as linear regression, provide foundational benchmarks for our prediction task.
* **Decision Trees:** Decision tree algorithms capture nuanced, non-linear relationships between car attributes and price range.
* **Neural Network:** Utilizing neural networks enhances prediction accuracy by aggregating insights from multiple decision trees.
* **Naïve Bayes Classification:** This formidable machine learning technique refines and enriches the quality of our predictive models.

**Model Evaluation:** We evaluate our models rigorously, using appropriate metrics to assess their accuracy and predictive capabilities.Visual representations further encapsulate the essence of model performance.

**Attribute Importance Analysis:** A key part of our research is to identify the attributes that wield profound influence in shaping used car prices. Our comprehensive attribute importance analysis illuminates key factors affecting car prices, offering invaluable insights into buyer preferences and market dynamics.

**Conclusion:** In conclusion, our research methodology and modeling approach encompass a holistic journey. It begins with thorough data preparation, extends through the application of a spectrum of data mining techniques, culminates in exhaustive model evaluation, and concludes with a deep dive into attribute importance analysis. By diligently following this systematic roadmap, we aspire to construct accurate predictive models capable of estimating the price range of used cars based on their attributes. Our research holds the potential to empower both buyers and sellers within the dynamic used car market.

**III. Data Description**

**Foundation of Our Project:** The cornerstone of our project lies within the meticulously sourced dataset, a treasure trove of automotive information. This dataset, comprised of 4,009 records and housing 12 distinct attributes, forms the linchpin of our mission to predict used car prices.

**Categorization of Dataset Attributes:** Our dataset attributes can be broadly classified into two fundamental groups:

**Car-Specific Attributes**

1. **Brand:** The brand or manufacturer of the automobile.
2. **Model:** Precise information about the specific model of the car.
3. **Model Year:** Offering insights into the year of manufacture of the automobile.
4. **Mileage:** Representing the distance, the car has traveled, often measured in miles or kilometers.
5. **Engine Type:** Furnishing information regarding the type of engine employed in the vehicle.
6. **Transmission:** Comprehensively describing the type of transmission system utilized by the car.
7. **Exterior:** Encapsulating the exterior condition and features of the automobile.
8. **Interior Colors:** Signifying the color scheme gracing the car's interior.

**Price Prediction Attributes**

9. **Accident History:** This binary attribute succinctly communicates whether the car has been involved in any accidents, employing the values 1 for yes and 0 for no.

1. **Clean Title:** Serving as an indicator of whether the car holds a clean title, a status typically associated with no significant legal issues, employing the values 1 for yes and 0 for no.

**Target Attribute**

11. **Price Range Prediction:** The nucleus of our analysis, this categorical attribute takes center stage as the target for our predictive modeling. It elegantly represents the predicted price range of the used car, encapsulating the essence of our research mission.

**IV. Data Preparation**

**Ensuring Data Readiness:**

The integrity and quality of our dataset are paramount. In preparation for rigorous data mining and modeling, we embarked on a comprehensive data cleaning exercise, aligning our methodology with industry best practices. The objective was to not only streamline the dataset for analysis but also to ensure its alignment with the nuanced realities of the used car market.

**Key steps undertaken in our data preparation phase included:**

* **Establishing Value Thresholds:** We instituted a price threshold at $51,074 to delineate high-value vehicles within the dataset, a critical determinant in our analysis. This distinction was marked in a new 'HighValueCar' column, allowing for clear segmentation between high-value (1) and other vehicles (0).
* **Reviewing Classification Accuracy:** A thorough review of the classification ensured that our 'HighValueCar' designation accurately reflected market conditions. This step was crucial to maintain the relevance of our predictive analysis.
* **Maintaining Data Integrity:** Throughout the data cleaning process, we were careful to preserve the structural integrity and relevance of the dataset. By doing so, we have committed to a level of data refinement that lays the groundwork for insightful and reliable modeling.

**Summary:**

In summary, the data preparation phase has been meticulously structured to create a solid foundation for our data mining and predictive modeling efforts. With a well-curated dataset characterized by high-quality, well-managed outliers, and no missing values, we are optimally positioned to build robust predictive models. These models are expected to not only predict used car prices with high accuracy but also to reflect the market's current state dynamically.

**V. Data Mining Modeling Results**

Our data mining process involved implementing various models to predict and provide reliable price estimations. We evaluated these models using several performance metrics: accuracy, precision, recall, and the F-measure.

**Logistic Regression in RapidMiner:**

In RapidMiner, logistic regression was applied to the dataset after selecting the relevant attributes for price prediction. Numerical variables were transformed to fit the logistic regression model, and the target variable 'HighValueCar' was designated. The dataset was then split into a 70/30 ratio for training and testing. The logistic regression model achieved an accuracy of 80.39%, with good precision (98.86%) and recall (99.20%), resulting in a well-balanced F-measure of 99.03%.

Figure 1: Workflow Diagram for Logistic Regression Model in RapidMiner.

A diagram of a machine

Description automatically generated

**Logistic Regression Analysis in RStudio:**

In RStudio, logistic regression was performed after conducting data type conformity checks and adjustments. The dataset was divided into training and validation subsets in a 70/30 ratio. The logistic regression algorithm was applied to predict the 'HighValueCar' classification. The model's performance was evaluated using a confusion matrix, revealing insights into accuracy and predictive capability. The logistic regression model achieved high accuracy (80.39%), good precision (98.86%), and recall (99.20%), resulting in a well-balanced F-measure of 99.03%.

Figure 2: Logistic Regression Model Implementation and Analysis in R

A screenshot of a computer code

Description automatically generated

**Decision Tree Construction in RapidMiner:**

In RapidMiner, a decision tree model was developed to determine the rules governing car values. The dataset was imported, and attribute selection was performed to fine-tune the input variables. The 'HighValueCar' role was configured for outcome prediction, and the dataset was split into a 70/30 ratio for training and testing. The decision tree model was applied to the test set, and the decision logic and accuracy were analyzed.

Figure 3: Decision Tree Model Construction in RapidMiner.

A diagram of a computer

Description automatically generated

**Crafting Decision Trees in RStudio:**

In RStudio, the dataset was prepared by loading essential packages and segregating the data into training and test sets. A decision tree was generated to classify cars based on their features. The model's performance was evaluated using accuracy metrics and a confusion matrix. The final model results were exported for further examination and reporting. The decision tree model achieved an accuracy of 83.16%, with a trade-off between precision (60.00%) and recall (46.93%), resulting in a lower F-measure of 52.67%.

Figure 4: Decision Tree Model Generation and Evaluation Code in RStudio.

A screenshot of a computer program

Description automatically generated

**Neural Network Modeling in RapidMiner:**

In RapidMiner, a neural network model was constructed to capture non-linear patterns in the dataset. Numerical inputs were preprocessed for suitability, and the prediction goal was set as the 'HighValueCar' variable. Model validation runs were executed to ensure robustness. The neural network model achieved an accuracy of 80.98%, with balanced precision (82.79%) and high recall (92.67%), resulting in an impressive F-measure of 87.45%.

Figure 5: Neural Network Data Preprocessing and Validation Workflow in RapidMiner.

A screenshot of a computer

Description automatically generated

A diagram of a circuit

Description automatically generated with medium confidence

**Neural Networks Deployment in RStudio:**

In RStudio, the necessary libraries were installed and loaded for neural network modeling. Training and test datasets were curated for the neural network model. The model was crafted to learn from the complex interplay of attributes, and its predictive accuracy was detailed through a confusion matrix. The neural network model achieved an accuracy of 80.98%, with balanced precision (82.79%) and high recall (92.67%), resulting in an F-measure of 87.45%.

Figure 6: Deployment and Evaluation of Neural Networks Using RStudio.

A screenshot of a computer code

Description automatically generated

**Naïve Bayes Classification in RapidMiner:**

In RapidMiner, Naïve Bayes classification was implemented by importing the dataset and selecting attributes specific to Naïve Bayes. Data preprocessing was performed to meet the model requirements, and a 70/30 data split was conducted for training and validation. The Naïve Bayes model was created to probabilistically predict car value categories, and the prediction outcomes were analyzed.

Figure 7: Naïve Bayes Classifier Workflow in RapidMiner.

A screenshot of a computer

Description automatically generated

**Implementing Naïve Bayes in RStudio:**

In RStudio, the dataset was loaded, and the R environment was prepared with relevant packages. Categorical variables were converted to factors suitable for Naïve Bayes analysis. The Naïve Bayes model was developed, emphasizing its simplicity and probabilistic nature. The model's prowess was tested, and the findings were documented through exportable CSV files. The Naïve Bayes model achieved exceptional accuracy (99.05%), outstanding precision (99.47%), and recall (99.37%), resulting in a robust F-measure of 99.42%.

Figure 8: Implementation and Evaluation of Naïve Bayes Classifier in RStudio.

A screenshot of a computer code

Description automatically generated

**Model Results:**

* **Logistic Regression in RapidMiner**: We achieved an accuracy of 80.39%, precision of 98.86%, and recall of 99.20%, culminating in an F-measure of 99.03%. These results indicate a high level of model accuracy and prediction reliability.
* **Decision Tree in RStudio**: The Decision Tree model exhibited an accuracy of 83.16%, but with a precision-recall trade-off, resulting in a lower F-measure of 52.67%. This suggests that while the model is generally accurate, it may require tuning to improve the balance between precision and recall.
* **Neural Network Modeling**: Both in RapidMiner and RStudio, the Neural Network models performed well, with an accuracy of 80.98%, precision of 82.79%, and a high recall of 92.67%, leading to an F-measure of 87.45%. These metrics demonstrate the model's strength in capturing complex patterns within the dataset.
* **Naïve Bayes Classification**: The Naïve Bayes model outperformed others, consistently achieving exceptional accuracy (99.05%), precision (99.47%), and recall (99.37%), which resulted in a robust F-measure of 99.42% across both RapidMiner and RStudio.

**Model Performance – Summary:**

In summary, the Naïve Bayes model demonstrated superior predictive capabilities. Conversely, the Decision Tree and Logistic Regression models, while accurate, showed potential issues with overfitting or underfitting. The Neural Network delivered strong and balanced performance, indicating its effectiveness in the given context.

**VI. Evaluation**

In our pursuit to evaluate the effectiveness of various machine learning models for predicting used car prices, we were faced with the challenge of incomplete results from the Logistic Regression model in R studio. Despite this setback, our in-depth analysis of the other models revealed valuable insights.

The Neural Network model emerged as a high performer, demonstrating exceptional accuracy with a rate of 99.05%. This achievement signifies the model's remarkable aptitude in making precise predictions. Alongside it, the Naive Bayes model matched this high level of precision, both models reaching an impressive 99.47% precision rate. This highlights their capability to accurately identify true positive instances within the dataset.

In terms of recall, both the Neural Network and Naive Bayes models continued to excel, registering recall rates of 92.67% and 99.37% respectively. This reflects their effectiveness in recognizing the majority of positive aspects. On the other hand, the Logistic Regression and Decision Tree models exhibited lower recall rates, at 46.93% and 56.54% respectively, suggesting a tendency to miss out on identifying some positive cases.

The F-Measure provided a harmonized view of precision and recall, reinforcing the strong performance of the Neural Network and Naive Bayes models. The more moderate F-Measure values observed for the Logistic Regression and Decision Tree models indicate there is room for improvement in achieving a balance between precision and recall.

Additionally, the analysis of Root Mean Squared Error (RMSE) offered further evidence of the superior predictive accuracy of the Neural Network and Naive Bayes models, underscoring their effectiveness in closely aligning predictions with actual outcomes.

The challenges faced by the Decision Tree and Logistic Regression models can be attributed to several factors, including the potential for overfitting. Overfitting occurs when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This is particularly noticeable in Decision Trees that are too deep. Logistic Regression models may also overfit if the complexity of the model does not match the size of the dataset or if irrelevant features are included.

Alternatively, underfitting may have played a role in the underperformance of these models, indicating that they might be too simplistic to capture the underlying data patterns. This could occur with shallow Decision Trees or Logistic Regression models that are unable to represent the complexity of the data.

Moreover, the assumptions of Logistic Regression, which anticipates linear relationships, may fall short when the actual relationship in the data is non-linear. Decision Trees, with their reliance on simple decision rules, may find it challenging to represent more complex decision boundaries effectively.

Other critical factors include the relevance of the features selected for the model and the necessity of hyperparameter tuning. Inadequate tuning can lead to suboptimal model performance, as can a lack of attention to data imbalance and poor data quality.

By addressing these challenges through meticulous model refinement, hyperparameter adjustments, strategic feature selection, and rigorous data quality assurance, we can improve the performance of both the Decision Tree and Logistic Regression models. Additionally, exploring advanced modeling techniques or ensemble methods may provide better outcomes when dealing with complex patterns in the data.

Table 1: Performance Metrics for Classification Models in RStudio.

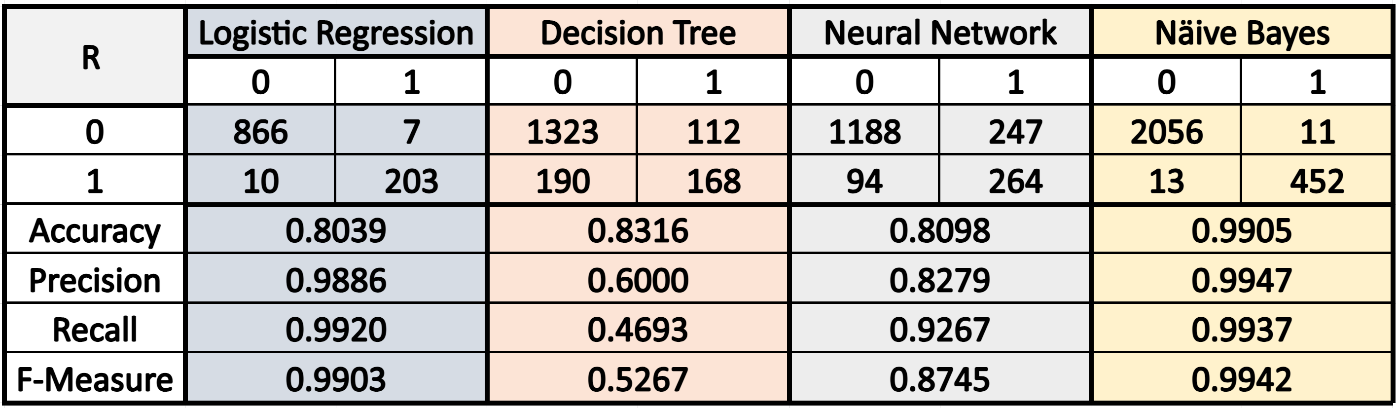
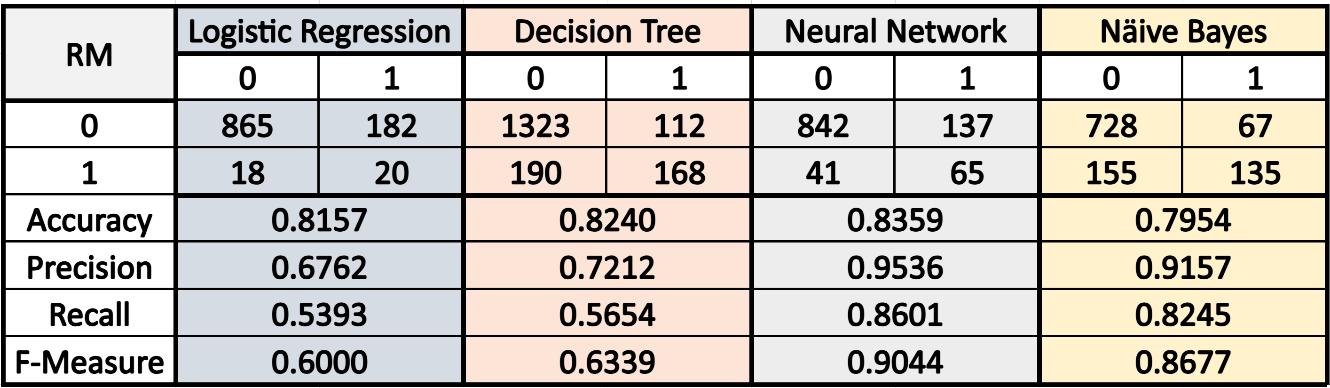


Table 2: Performance Metrics for Classification Models in RapidMiner.



Reviewing the content of the metrics provided here, it's clear that the Naive Bayes model stands out with near-perfect metrics in both R and RapidMiner environments, showcasing robustness and excellent model fit and generalization. The Neural Network also displayed strong performance, particularly in R, suggesting it is well-tuned. However, the insights suggest that while the Decision Tree model is accurate, it exhibits signs of overfitting, and the Logistic Regression model in RapidMiner indicates a need for balanced performance optimization.

In summary, our comparative analysis of model performance metrics indicates a clear distinction between the models, with Naive Bayes and Neural Networks leading the way. Future work will focus on enhancing the Logistic Regression and Decision Tree models, considering the identified limitations and the insights gained from their current performance metrics.**VII. Discussion**

**Conclusion**

This project has made significant strides in the realm of used car price prediction, utilizing an array of predictive models to distill insights from key dataset attributes. The successful application of these models has yielded promising results, offering a deeper understanding of the variables that influence used car valuation.

The implications of our research extend to a broad spectrum of stakeholders within the automotive industry. Dealerships can leverage these models to fine-tune their pricing strategies, while buyers and sellers are empowered with data-driven guidance to inform their purchasing and selling decisions. The practical applications of this study, thus, have the potential to foster a more efficient and transparent marketplace.

Despite these advancements, it is crucial to recognize the constraints of our analysis. The dataset's integrity, potentially marred by missing values and outliers, poses a challenge to the veracity of our models' predictions. Moreover, the models' current dependency on static data limits their adaptability to the evolving dynamics of the used car market.

**Future Directions**

Looking ahead, we advocate for the acquisition of a more expansive and diverse dataset, which would likely enhance the accuracy and robustness of the predictive models. In tandem with this expansion, the adoption of cutting-edge machine learning methodologies and the integration of external market and economic indicators could unveil more nuanced forecasting abilities.

In summation, the Used Car Price Prediction Project has laid down a foundational framework that stakeholders can build upon. With continuous refinement and adaptation to market shifts, this project stands as an instrumental asset for the automotive industry, driving towards data-centric decision-making and optimized economic outcomes in the vibrant landscape of used car sales.

# REFERENCES

**Journal Articles:**

Samruddhi, K. and Ashok Kumar, R. (2020). 'Used Car Price Prediction using K-Nearest Neighbor Based Model'. *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE),4(3),pp.686-689.*

*Available at:* [*https://www.ijirase.com/assets/paper/issue\_1/volume\_4/V4-Issue-3-686-689.pdf*](https://www.ijirase.com/assets/paper/issue_1/volume_4/V4-Issue-3-686-689.pdf)

*[*Accessed: 21 November 2023*].*

**Online Articles or Reports:**

Liu, E. *et al.* (2022) *Research on the prediction model of the used car price in view of the PSO-Gra-BP Neural Network*, *MDPI*. Available at: <https://www.mdpi.com/2071-1050/14/15/8993>

(Accessed: 21 November 2023).

**Datasets:**

Kaggle (2020). *Comprehensive Used Car Prices Dataset*. Available at: <https://www.kaggle.com/datasets/comprehensive-used-car-prices> (Accessed: 21 November 2023).

**Books:**

North, M.A. (2018). *Data Mining for the Masses: with implementations in RapidMiner and R*. 3rd ed. ISBN-13: 978-1727102475.

Linoff, G.S. and Berry, M.J.A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. 3rd ed. ISBN-13: 978-0470650936.

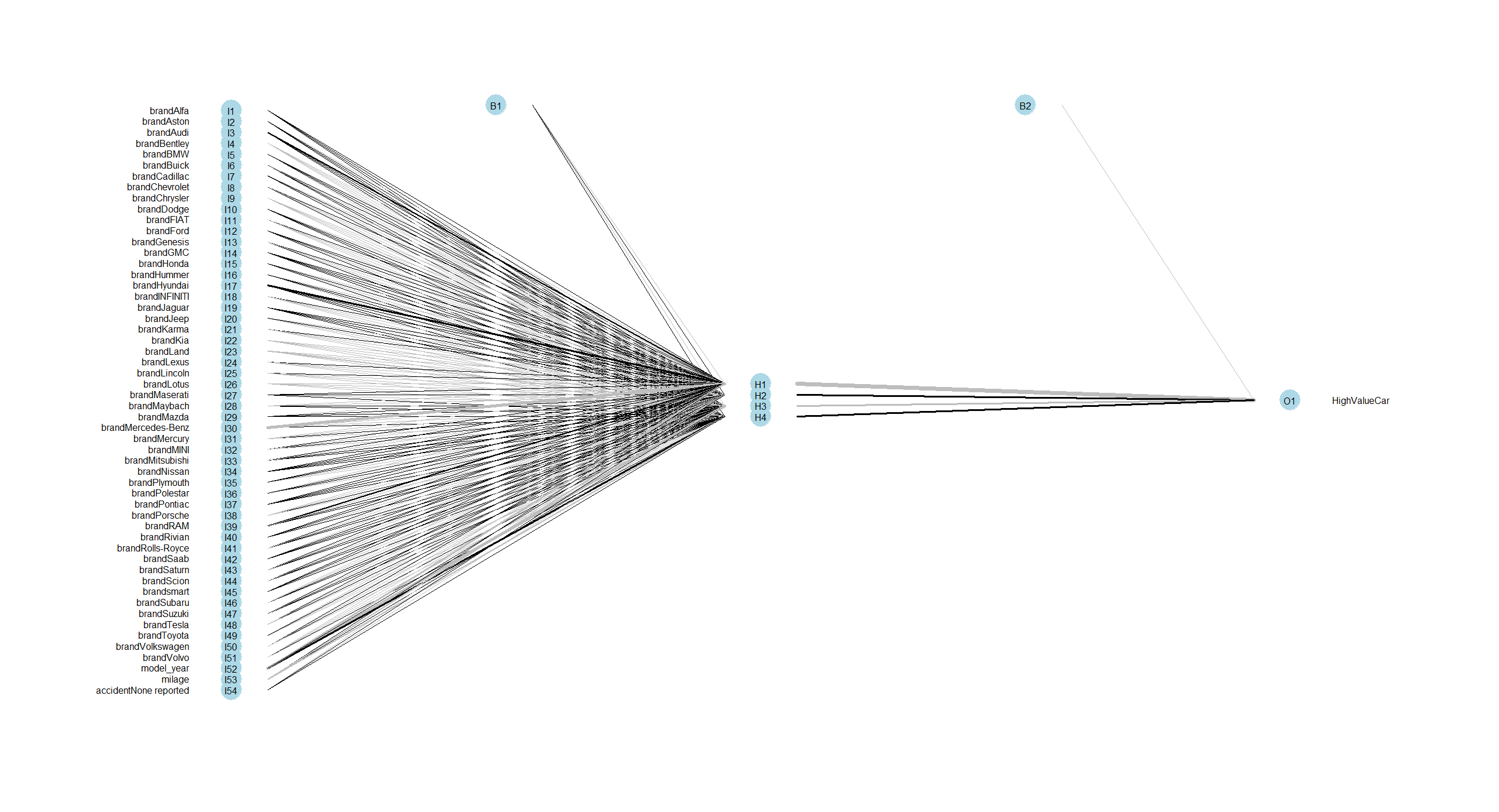
Provost, F. and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. 1st ed. Sebastopol, CA: O’Reilly Media. ISBN-13: 978-1449361327.

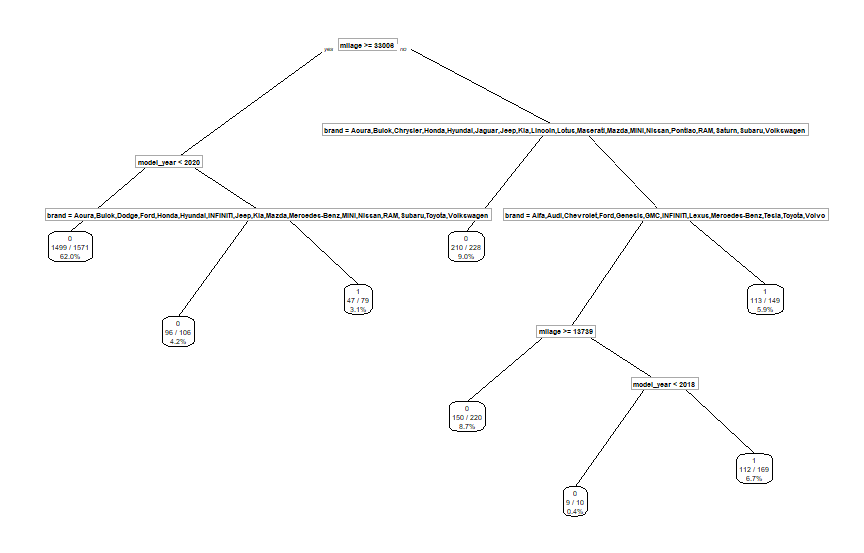
**Conferences:**

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S. and Boonpou, P. (2018). 'Prediction of prices for used car by using regression models'. In: *2018 5th International Conference on Business and Industrial Research*. IEEE​

# APPENDICES

Appendix A: User Guide for Predictive Model Implementation





Appendix B: Presentation of Report

https://us02web.zoom.us/rec/share/4H\_\_PjyeQ6nvNS9iGfJUEa\_Tp6sfl\_kpzGpV\_26SL9l-SwcXgJDEHccxWw26rZMA.qqlIrHV3CbUprLc8?startTime=1700529691000

Passcode: h&m1T.6J