DS 450-01

Predictive Analytics Project Proposal

Project Title: Predicting the Sale Price of Used BMW Cars

Jardin Dantzler

jdantzler@bellarmine.edu

January 16th, 2025

Executive Summary

The used car market is an essential segment of the automotive industry, with millions of vehicles sold annually. Determining the price of a used car is a multifaceted challenge, involving physical features, brand perception, and market trends. This project aims to develop a predictive model to estimate the sale price of pre-owned BMW cars using a dataset containing approximately 5,000 entries and 18 variables. By analyzing factors such as mileage, engine power, color, and car type, we will identify the most influential features affecting car prices. Using advanced machine learning models such as Decision Trees, Random Forests, and Gradient Boosting Machines, we will compare their performance and accuracy in predicting prices. This project will provide valuable insights for car dealerships, buyers, and sellers to make informed decisions in the used car market.

Project Idea

The goal of this project is to predict the sale price of used BMW cars based on various physical, technical, and categorical attributes. The analysis will focus on identifying key factors that significantly influence the price and developing a robust predictive model to accurately estimate car prices. By leveraging a well-structured dataset, we aim to provide actionable insights into the pricing dynamics of pre-owned BMW vehicles.

Background

Evaluating the price of used cars is a critical challenge for the automotive industry, as the value of a vehicle depends on numerous variables. Existing solutions often rely on generalized pricing algorithms or manual appraisals, which can be inconsistent and subjective. By developing a data-driven model, we can overcome these limitations and ensure consistent pricing.

The dataset used for this project contains detailed information on used BMW cars, including make, model, mileage, engine power, fuel type, and sale price. For this analysis, we will focus on Diesel and Petrol fuel types, as cars with other fuel types are underrepresented and considered outliers. The sale price, the target variable, will be modeled using various predictive analytics techniques. Additional features such as color, car type, and registration date will be analyzed for their influence on pricing trends.

Modeling

To predict the sale price of used BMW cars, we propose the use of the following predictive models:

1. **Decision Tree Regressor**: This model is simple yet effective for capturing non-linear relationships in data and interpreting feature importance.
2. **Random Forest Regressor**: An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
3. **Gradient Boosting Machines (GBM)**: A powerful algorithm that builds trees sequentially to minimize prediction errors, often outperforming other models in structured data tasks.

To evaluate the models, we will use metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Linear regression may also be implemented as a baseline for comparison.

Tools

The tools and technologies to be used in this project include:

- **Programming Language**: Python, R
- **Libraries**: Pandas (data manipulation), NumPy (numerical operations), Scikit-learn (machine learning), Matplotlib/Seaborn (visualizations), Caret Library
- **Data Visualization Tools**: Power BI or Tableau for creating interactive dashboards outside of Python or R

- **Jupyter Notebook/ Rstudio**: For exploratory data analysis and model development

These tools were selected for their robustness, ease of use, and industry-wide adoption for predictive analytics projects.

Conclusion

This project aims to leverage predictive analytics to create a reliable model for estimating the sale price of used BMW cars. By identifying the key factors influencing pricing and evaluating multiple predictive models, we aim to provide insights and tools that can enhance decision-making in the used car market. Successful implementation of this project will benefit stakeholders by improving pricing transparency and accuracy.

References

1. Kaggle BMW Pricing Challenge Dataset: https://www.kaggle.com/danielkyrka/bmw-pricing-challenge
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
3. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.