Joshua Grant

12/04/2024

CIS 390 – Supervised ML

**Business Use Case: Synopsis**

***Background***

A motorcycle business that we provided insights for buys used motorcycles to sell to their clients The company also offered a maintenance warranty for the first 100 days after purchase. They also wanted to gain the most knowledge of their products. This motorcycle company asked us to build a machine learning model that can **predict motorcycle resale prices**. To achieve this, we demonstrated the CRISP-DM process to examine the data and select an appropriate model for offering insights into this use case.

***Use Case Overview***

Motorcycles are a popular mode of transportation in many parts of the world. Their resale value is influenced by a variety of factors, including age, mileage, engine size, brand, and more. For this motorcycle business, they wanted to understand the determinants of resale prices to be able to maximize profits and improve customer satisfaction. The goal of this project was to leverage supervised machine learning to build a predictive model that can estimate the resale price of motorcycles based on various features.

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) process, utilizing all six stages to our advantage. These stages – Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment – guide the systematic development of the model.

***Key Insights***

For the **“Business Understanding”** stage of the CRISP-DM, we need to ensure that we understand the business problem, business objective, and key goals of the project. The problem is that the motorcycle resale market lacks transparency and consistency, as prices can vary significantly depending on various factors. This could lead to sellers overpricing or underpricing their motorcycles, leading to inefficiencies in the market. The business objective is simple: maximize profitability by being able to accurately predict the resale prices of used motorcycles. By using predictive data analytics, the business should be able to set optimal pricing which will improve customer satisfaction and, of course, increase profits. The key goals for this project are to develop an accurate price prediction model using supervised machine learning techniques.

During the **“Data Understanding”** stage of the CRISP-DM cycle for this project, we made some key discoveries in our dataset. It contains features such as the **selling price** (target variable), **year**, **kilometers driven**, **ex-showroom price**, **seller type**, and **owner**. We also noticed a significant number of missing values in the **ex-showroom price** column. These missing values were handled in the **“Data Preparation”** stage for this project. The dataset contained both categorical (name, seller type, owner) and numerical (year, selling price, kilometers driven, ex-showroom price) type features. Some outliers were also detected in the dataset. For example, in the **selling price** feature, there were some values significantly higher than others. This was likely due to the motorcycle being some sort of “limited edition” or historical vehicle. These outliers were also handled later.

The **“Data Preparation”** stage of the CRISP-DM is when we actually handled and dealt with all our findings from the previous stage. We first cleaned the dataset, imputing missing values in the **ex-showroom price** column with the median. We also removed outliers using the interquartile range for each feature. Next, we performed feature engineering, creating an **age** feature, a **usage intensity** feature, and a **brand** feature. These features ended up being instrumental in the model’s performance. We also encoded categorical variables to be numerical to comply with the machine learning algorithms used later, as well as scaled numeric features which made them more suitable for regression models sensitive to feature scales. The results of these transformations were then saved to a new dataset, which was used for the rest of this project.

Up next was **“Modeling”** and **“Evaluation”**. These both went hand in hand together. Two different algorithms were tested during this stage: linear regression and random forest regression. We used linear regression to get a baseline model, showing decent performance. It had a large value for mean squared error, indicating high variability in prediction errors, but the mean absolute error was able to tell us that we were only off by around $12,000 on average, which isn’t bad considering the prices of these motorcycles in the dataset can get over $200,000. Visualizations were then produced that were able to show us that our model was successfully capturing a relationship between features, but that it consistently predicted a selling price higher than the actual selling price. We then built a random forest regressor, which showed much more promising results. The r ² value was almost 30% higher, meaning there was a 30% increase in variance in the target variable being explained by the model. The mean absolute error also indicated better performance, telling us that the average miss was around $8,740, quite the improvement from linear regression. After this, hyperparameter tuning was applied using a grid search. After the optimal parameters were applied, the model was retrained and once again showed improvement across the board, dropping the average miss to about $8,600. Visualizations were once again produced to tell us more about the model. We discovered that the **brand** and **age** feature that we engineered earlier, were two of the three most important features to the model’s success. We also found that we were able to eliminate the tendency of having too high of a guess that we saw in the linear regression model.

This leads us to **“Deployment”**. Now that the final model has been built, it is ready for deployment in a business context, enabling the company to input motorcycle features and instantly receive a predicted resale price. Potential applications include dynamic pricing strategies, trade-in value estimation, and personalized customer offers.

***Conclusion***

This project demonstrated the power of supervised machine learning in addressing real-world problems. By identifying the key drivers of a motorcycle’s resale price, we were able to build an accurate predictive model. This solution adds value for both the company and customers and should be able to improve customer satisfaction and profits. Insights from the model can inform pricing strategies and improve company decision making, making this model exactly what the motorcycle company needed.