Joshua Grant

09/18/2024

CIS 390 – Supervised ML

**Business Use Case: Business and Data Understanding**

For this assignment, I’ll be selecting the first use case provided in the course content: Predicting Motorcycle Resale Prices. The objective of this use case is to build a model that predicts the resale price of a motorcycle based on features such as age, mileage, engine size, and brand. To accomplish this goal, I will be using the CRISP-DM process (Cross Industry Standard Process for Data Mining), starting with the first two phases: Business Understanding and Data Understanding.

***Business Understanding:***

Business Problem: The motorcycle resale market lacks transparency/consistency and prices can vary significantly depending on various factors including brand, condition, mileage, and more. Sellers may overprice or underprice their motorcycles, leading to inefficiencies and unsatisfactions in the market. The motorcycle business I am helping is looking for a model to help expedite this process and make it more reliable, accurate, and efficient.

Business Objective: The motorcycle business aims to maximize profitability by being able to accurately predict the resale prices of used motorcycles. The company buys used motorcycles and offers a 100-day maintenance warranty. By using predictive data analytics, the business should be able to set optimal pricing for the resale of motorcycles, improving customer satisfaction, sales margins, and of course, profits.

Key Goals: I want to develop accurate price predictions using supervised machine learning techniques. This model should be able to predict the resale value of motorcycles based on key attributes such as age, mileage, engine size, and brand. The business should be able to use insights from this resale price model to optimize purchasing decisions for used motorcycles. This will ensure that the company acquires motorcycles that are more likely to have a higher resale value. Since this model will help the business to offer more competitive and more accurate prices, the business will be able to attract more customers and build long-term relationships through this reliable pricing model.

***Data Understanding:***

Dataset Description: The dataset contains information on motorcycles listed for resale, with key attributes such as the name, selling price, year of manufacture, seller type, ownership status, kilometers driven, and ex-showroom price. There are some missing values in the ex-showroom price column, denoted by a “?”, which could impact the analysis.

Attribute Descriptions:

* **Name:** A categorical value representing the make and model of the motorcycle. This can provide important brand and model information, which will significantly influence the resale value.
* **Selling Price:** A numeric feature representing the price at which the motorcycle was listed for resale. This is the target feature that we aim to predict with our model.
* **Year:** A numeric value representing the year the motorcycle was manufactured. Newer models generally have higher resale prices, so this will be an important attribute.
* **Seller Type:** A categorical value representing whether the seller is an individual or dealer. Dealer listings could have different pricing patterns compared to individual sellers.
* **Owner:** An ordinal value describing whether the motorcycle is being sold by the first owner, second owner, etc. The number of owners typically affects resale value.
* **Kilometers Driven:** A numeric feature describing the total kilometers driven by the motorcycle. Higher milage will lead to lower resale value.
* **Ex-Showroom Price:** A numeric value that gives the price at which the motorcycle was originally sold for in the showroom. This feature is available for some motorcycles, but not all, and can be crucial for determining the depreciation rate over time.

Initial Observations: We can see that there are missing values in the ex-showroom price column. These missing values may need to be addressed through imputation, as upon first glance it seems to be too many to remove. The dataset contains both categorical (name, seller type, owner) and numerical type features (year, selling price, kilometers driven, ex-showroom price), which will also need to be processed accordingly. Some motorcycles also seem to have extreme values (very low or very high prices). These outliers should be explored further to see how many there are and to see if they are just simply outliers or can tell us something about the data.

Plans for Exploratory Data Analysis: We can analyze the correlation between numerical variables like kilometers driven, year, ex-showroom price, and selling price. This will help to identify which features have the strongest impact on the resale price. We can also look at the distributions related to some of the categorical features such as seller type and ownership. By analyzing how the resale price varies between dealers and individual sellers, as well as between first and subsequent owners, we can provide insights on pricing patterns.