Joshua Grant

09/26/2024

CIS 390 – Supervised ML

**Business Use Case: Data Preparation**

For this assignment, I’ll be laying out my plans for creating the ABT (Analytical Base Table) and discussing my plans for data preparation for my project. As a reminder, my use case is 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 provided in the data.

***Plan for the ABT:***

1. Data Collection:

The data will be collected through the dataset provided to us on Blackboard. This dataset has a variety of features that we’ll use to build our ABT, and eventually, our model. These features we have access to are the motorcycle name, the selling price, the year of the bike, the seller type, the type of owner (1st owner, 2nd owner, etc.), kilometers driven, and the showroom price. We also may want to see if we have access to any external data, such as engine size, market demand, regional prices, and others.

1. Data Cleaning:

We’ll have to check for outliers in our data, handle missing values, and correct inconsistencies we find. For example, the showroom price in our dataset is missing in a large percentage of the datapoints, so this will have to be addressed, either by data imputing or removal. Some showroom prices and selling prices also could be deemed as outliers. We will also have to see if any of our features need to be standardized, and we will ensure that all numerical fields are in the same unit system.

1. Feature Selection:

In this step, we’ll identify which variables will influence motorcycle resale prices. This includes continuous variables such as age, mileage, engine size (if available), selling price, and categorical variables such as brand, motorcycle type (if available), year, etc. We will also drop irrelevant columns/features during this step.

1. Feature Engineering:

This is where we will create new features from existing data. Some of these could be the age at initial sale, which can be derived from subtracting the manufacturing year by the sale year. Another could be usage intensity, which could be achieved by using the mileage and age features. We can also encode categorical variables by converting features such as brand into numerical formats using techniques such as one-hot encoding. We will also normalize and scale any numerical features that require it to ensure all variables are on a comparable scale.

1. Preliminary Data Analysis:

In this step, we will run exploratory data analysis on the dataset to understand patterns and correlations. We will also utilize creating visualization to help us better understand the data and how its features are related. This could include creating histograms, scatterplots, bar charts, and correlation matrices.

1. Creating the ABT:

This will be the part where we actually build our analytical base table. We will define the target variable (resale price) and ensure that all predictor variables are included and formatted appropriately. We will also split the data into training and testing sets, ensuring that the ABT contains all relevant features for modeling.