Project 3 – Milestone 1

Topic –

This project will look at attributes from previous BMW Car Sales and will aim to build a predictive model that can accurately predict whether a BMV vehicle listing will result in a successful sale. By analyzing historical sales data, this project will identify key vehicle features that influence buyers purchasing decisions.

Business Problem –

Right now the car market, new and used, is very competitive. Car sellers need insights into what features influence the sale of a vehicle. Knowing which attribute sell the best will allow them to market these attributes when showing cars and help them to spend less time on attributes that contribute less the appeal of the car.

This project will answer the following question: 'What characteristics best predict if a BMW car will sell?'. The goal is aid dealerships and sellers by identifying key factors that boost sale potential.

Datasets –

This data set was accessed through Kaggle and includes information about BMW Car Sales. It includes entries with key factors with details about each car sold including; model, year, milage, transmission type, engine size, fuel type, and sale status. This data form the basis for training and evaluating classification models.

Methods –

I will apply the following data analysis modeling techniques for this analysis:

Exploratory Data Analysis (EDA) – This will include summary statistics and visualizations to help understand the distributions, identify trends, and detect any possible anomalies, missing values or outliers.

Data Processing – This will clean up the data to ensure the results are as accurate as possible. This will include; cleaning missing entries (either dropping them or filling them), encoding categorical variables for easier use in the models. Feature engineering may also be necessary to create variables from the existing data that work better with the models.

Predictive Modeling – This analysis will utilize predictive modeling techniques such as, Logistic Regression, Random Forest and XG Boost. Logistic Regression will estimate the probability of a binary outcome, if the car is sold or not, based on input features. The Random Forest Model will build decision tress on subsets of the data and average their

outputs in order to model complex interactions between features. And the XGBoost Classifier will help increase the accuracy of the results.

Model Evaluation – In order to evaluate the accuracy of the models, I will use accuracy, precision, F1-score, and confusion matrices. Cross-validation techniques may also be employed to ensure the model is strong.

Ethical Considerations –

There are some ethical considerations that should be considered when conducting this analysis. First, it should be noted that there may be bias in historical data. If certain types of cars are overrepresented, the model might unfairly favor them in predictions. Another is the transparency of the model to the sellers. Sellers may not trust a model or predictions made if they do not understand the reason/ explanation for the results. The sellers need to not only understand what the results mean, but how they were formulated. The last is privacy, while the dataset does not contain personal identifying information about the buyers, care must still be taken in how insights are generalized or interpreted.

Challenges/ Issues –

There are a few challenges that could impact the project and should be considered. It is possible that we could see a class imbalance with some attributes. For example, if there a significantly more "sold" listings than "not sold" listing the model performance could be skewed. We could also end up seeing multicollinearity, where some features maybe highly correlated. This could also affect model interpretation. There is also a chance of model generalization. It is important that the trained model performs well on unseen data and ensuring this will require careful validation.

References –

Chen, T., & Guestrin, C. (n.d.). *XGBoost Documentation*. Retrieved July 22, 2025, from https://xgboost.readthedocs.io/en/stable/

Junaid512. (2023). *BMW Car Sales Classification Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/junaid512/bmw-car-sales-classification-dataset