Project 3 – Milestone 2

Alyssa Graham

DSC 680

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.

Background/ History –

BMW is a globally recognize auto manufacturer that is known in the car industry for engineering and innovation. Its vehicle listings (both new and used) are posted across online platforms with varying success in sales. Historically, certain combinations of features such as, model, transmission type, engine size, etc.) have led to higher sales performance. Using these machine learning techniques, this project uses historical sales data to uncover these patterns, siding dealers and marketers in improving sales rates.

Data Explanation –

This dataset comes from Kaggle and includes 50,000 BMW vehicle listings. These records cover data from many different variables including; model, fuel type, color, year, mileage, engine size, price, and sales classification.

Some steps were taken to prepare the data before the analysis. Label encoding was used on categorical features. Duplicate and empty values were removed. An EDA was conducted to inspect distributions and detect any possible outliers.

Methods –

The dataset was cleaned, using steps stated previously, before the analysis began to ensure the most accurate results. An Exploratory Data Analysis (EDA) was conducted to look at the summary statistics for the dataset. Visualizations from this EDA were used to identify trends that could influence future analysis.

Supervised machine learning was used to classify BMW car listings as having either a "High" or "Low" sales performance based on vehicle characteristics. Label encoding was

used to convert categorical variables into numerical variables. The target variable, sales_classification, was binary-encoded for classification. The data was then split into 80% training and 20% testing sets. Logistic Regression, Random Forest Classifier, and XGBoost Classifier were trained with the data.

Analysis –

A correlation matrix was created to examine all the numeric features in the data set. This can help identify which variables are most relevant for predicting car sales performance and detecting multicollinearity. Figure 1 shows us the correlation matrix from this analysis. We see a very strong positive correlation between sales classification and sales volume. This was expected since sales volume is a direct measure of popularity. To avoid data leakage, sales volume was excluded from predictive modeling. This matrix also shows us that there is a weak positive correlation between Price and Sales classification. This suggests that pricing alone does not have a strong linear influence on sales classification. This makes sense because the price/ value will vary depending on other features. We see a weak negative correlation between Mileage and Sales Classification. Cars with higher mileage are more likely to be classified as low. This makes sense because cars with more mileage are generally less desirable. While this is not a complete guide to feature selection, this matrix provides an essential first step in understanding the data structure.

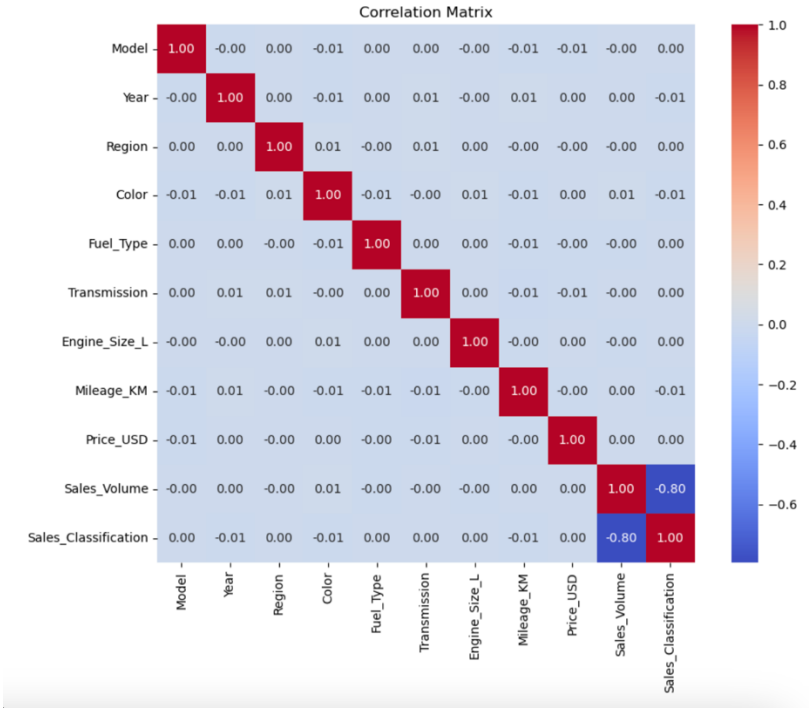


Figure 1: Correlation Matrix

Logistic Regression used a linear model as a baseline for binary classification. It estimates the probability that a given input belongs to a certain class. In this case, it helps determine if a car will be classified as “High” or “Low” sales. We see the results for the Logistic Regression of BMW Sales in Figure 2. The coefficients show the direction of each features effect. This model achieved and overall accuracy of 0.7, but it did not detect and “Low” sales listings. This model classified all test samples as “High” sales. These extremely imbalanced predictions indicate a biased boundary due to a class skew. With a Logistic Regression Model, relationships are assumed to be linear with likely oversimplifies real world auto sales data.

```
=== Logistic Regression ===
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      3032
           1       0.70      1.00      0.82      6968

    accuracy                           0.70     10000
   macro avg       0.35      0.50      0.41     10000
weighted avg       0.49      0.70      0.57     10000

=== Random Forest ===
              precision    recall  f1-score   support

           0       0.25      0.01      0.03      3032
           1       0.70      0.98      0.81      6968

    accuracy                           0.69     10000
   macro avg       0.47      0.50      0.42     10000
weighted avg       0.56      0.69      0.58     10000

=== XGBoost ===
              precision    recall  f1-score   support

           0       0.27      0.03      0.06      3032
           1       0.70      0.96      0.81      6968

    accuracy                           0.68     10000
   macro avg       0.49      0.50      0.43     10000
weighted avg       0.57      0.68      0.58     10000
```

Figure 2: Logistic Regression Results

In Figure 3, we see the results from the Random Forest Classifier Confusion Matrix. This model shows the complex relationships between features and ranks their importance. This model produced a better balance that the Logistic Regression Model, but we still see that the “High” sales class is still heavily favored. Random Forest received an accuracy score of 0.69. This model does a better job at handling the non-linearity of this data set. However, we still see a somewhat low recall for the “Low” sales listings.

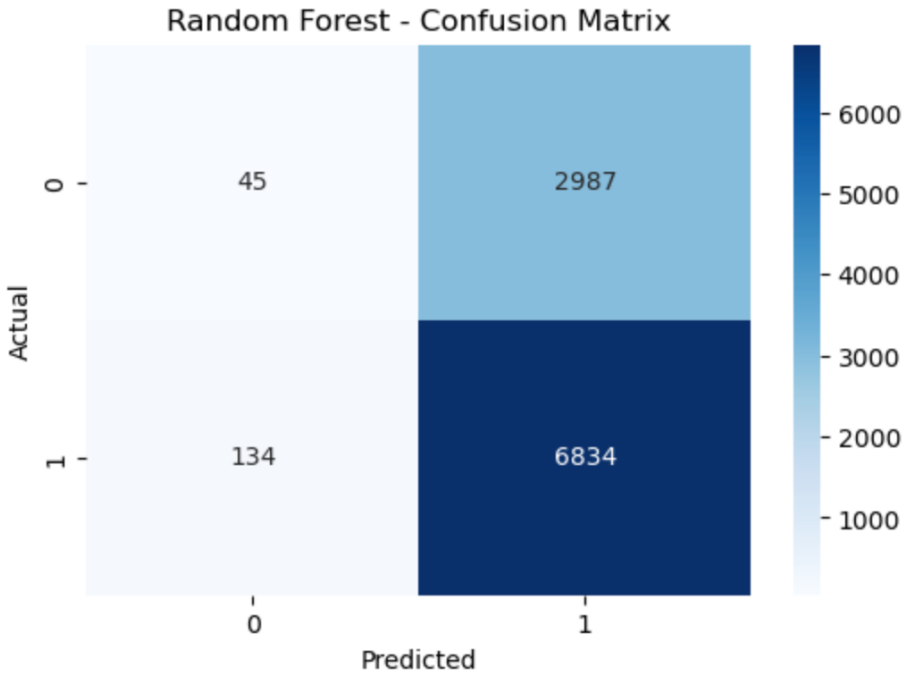


Figure 3: Random Forest Confusion Matrix

Figure 4, shows the features that have the most influence on car sales and how much they contribute to the sales. Based on this model, Price, Mileage, and Engine Size are top contributors to sales.

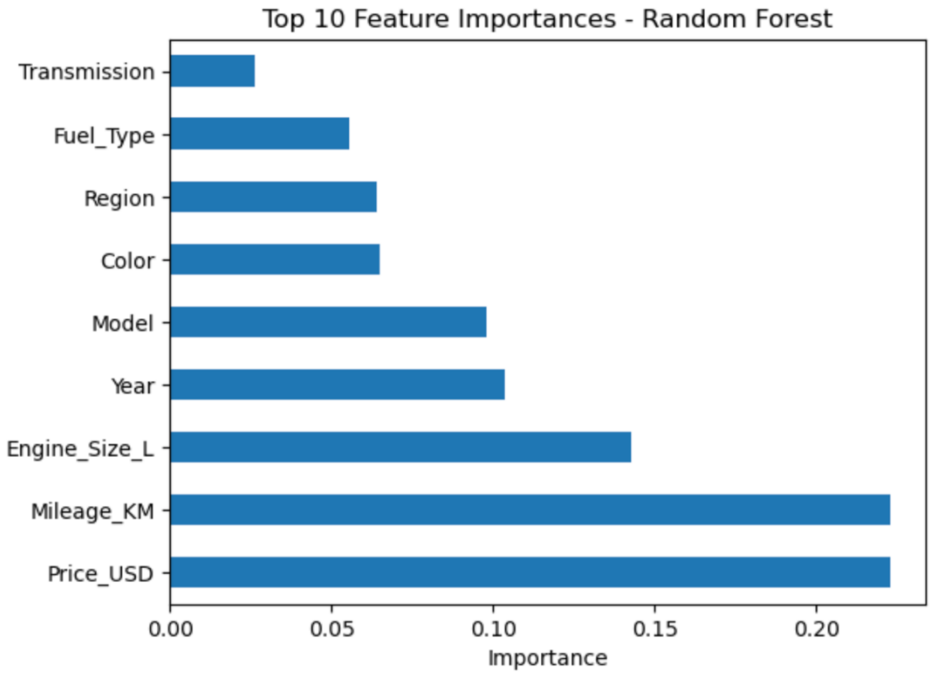
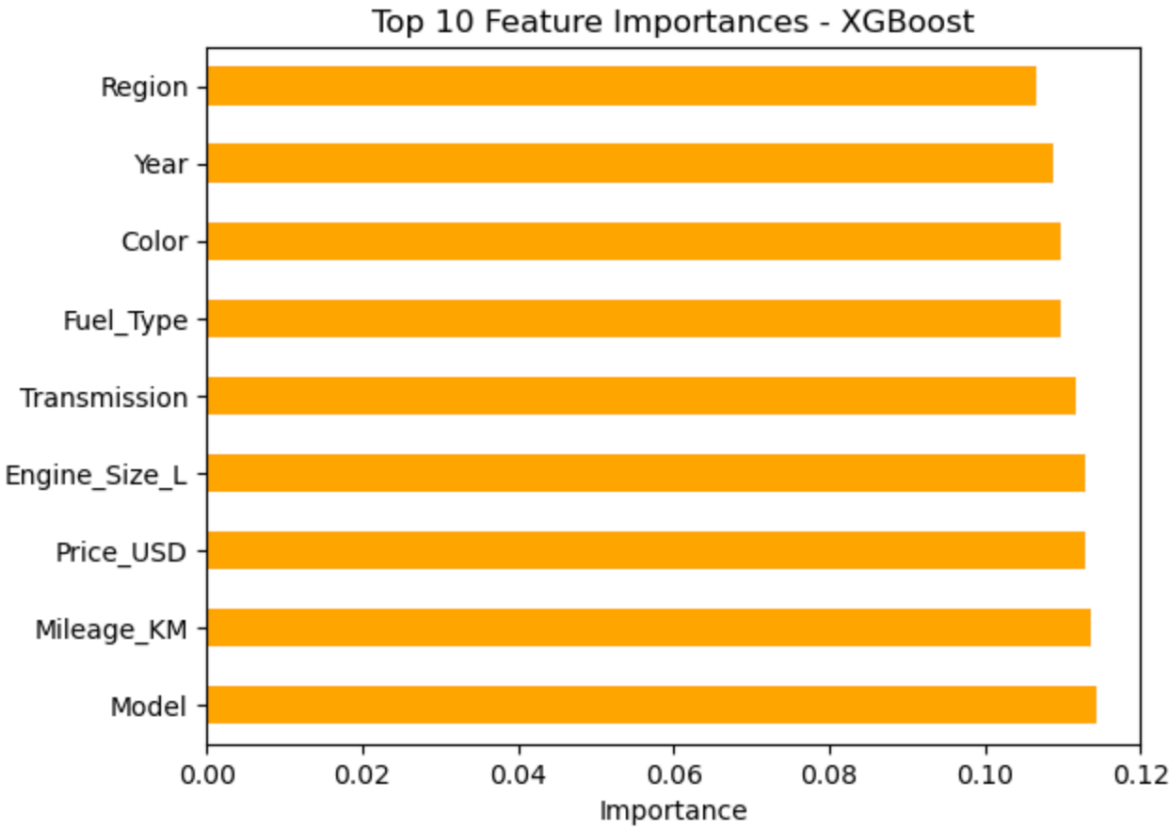


Figure 4: Random Forest Feature Importance

With the XGBoost Model, we see a similar performance to the Random Forest model with an accuracy score of 0.68. The XGBoost model had slightly better precision and recall for the “Low” class, but it was still low. Since the XGBoost model is a gradient boosting model, it was better able to learn from errors but it still shows bias toward the dominant class. This model helped to focus on harder to classify samples and supported feature importance output for better interpretability. In this model, all features have relatively equal importance. This suggests that the model is not overly reliant on any single feature. Instead it draws from a balanced combination of inputs. This may not be the best model to provide assistance with increasing car sales.



Conclusion –

This project applied classification models to predict BMW sales based on historical listing data. XGBoost and Random Forest outperformed the Logistic Regression. While these models did effectively identify “High” sales listings, they still struggled to correctly classify “Low” sales due to class imbalance. From these results, the Random Forest may be the most helpful to dealerships. From the Random Forest Model, we see that the key predicting features include price, mileage, and engine size. Although predictive accuracy was solid, improving class balance and interpretability may be helpful before deployment.

Assumptions –

The analysis assumes that the dataset from Kaggle accurately represents real-world BMW sales listings across diverse regions and that the sales classification labels reflect real commercial outcomes. We can also assume that historical patterns in features like price, engine size, and mileage continue to influence buyer behavior similarly today.

Limitations –

This dataset appears to be imbalanced with a majority of sales listings labeled as “High”, skewing predictions. It is also possible that some features may be proxies for others which can introduce some redundancy. This model also does not currently account for changes over time or trends in buyer preference.

Challenges –

There were a few challenges when running these models. First, all models showed extremely low recall for the minority “Low” class. Also, features like sales volume were initially strongly correlated with the label and had to be excluded.

Future Uses –

This information can help salesmen and BMW dealers determine what is most important to buyers when buying a BMW. This model could be adapted to other car brands or used car platforms to offer similar information to other car dealerships.

Recommendations –

Class rebalancing techniques could be added to improve the detection of low-performing listings. Hyperparameter tuning could be used to improve the model generalization across different data subsets.

Implementation Plan –

These models can be tuned using hyperparameters and them deployed to dealerships to be integrated into dealership management software. Predictions and outcomes of sales can be documented over time to detect changes in sales and new trends.

Ethical Assessment –

There is a risk of bias as models trained on historical data may perpetuate past preferences. This could potentially unfairly favor certain models, price points, or regions. It is also important to keep the privacy of the buyers. While the current data was made anonymous, any future deployment should ensure buyer confidentiality. Dealers and sellers must also use these results s guidance and not as absolute decisions.

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. Retrieved from https://www.kaggle.com/datasets/junaid512/bmw-car-sales-classification-dataset

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. In *Advances in Neural Information Processing Systems* (Vol. 30). https://arxiv.org/abs/1705.07874

Questions from the Audience –

1. How was the target variable “Sales_Classification” defined—what makes a sale "High" or "Low"?
2. How did you verify the quality and completeness of the dataset?
3. Were external variables like dealer reputation or seasonal factors considered in the model?
4. Did any surprising relationships emerge during EDA?
5. Did you detect any multicollinearity between numeric features? How did you address it?
6. Why did you choose Logistic Regression, Random Forest, and XGBoost for your models?
7. What techniques did you consider (or apply) to address class imbalance?
8. How do we interpret the feature importance values from Random Forest or XGBoost?
9. How can dealerships or private sellers use these predictions in real-world sales strategies?
10. How would you improve or expand this project if you had more time or data?