**Audience Questions & Answers**

**Questions from the Audience –**

1. **How was the target variable “Sales\_Classification” defined—what makes a sale "High" or "Low"?**

The “Sales\_Classification” target variable was derived from the “Price” column. If the value in the price column was above a certain threshold the Sales\_Classification was labeled as “High” if the value was below the same threshold the “Sales\_Classification” was labeled as “Low”. This creates a binary classification for predicting the likelihood of high-value sales.

1. **How did you verify the quality and completeness of the dataset?**

Checking the data quality included verifying there were not duplicate rows, missing values were handled, correct data types were confirmed, value ranges for numeric features were validated, and categorical fields were checked to ensure they matched the expected values. Summary statistics and null values counts were reviewed before modeling.

1. **Were external variables like dealer reputation or seasonal factors considered in the model?**

The dataset did not include external variables like dealer reputation, seasonal sales trends, or economic conditions. These could be valuable for improving the predictive power but would require integration with additional sources.

1. **Did any surprising relationships emerge during EDA?**

Yes! Some relationships emerged that were not necessarily obvious before conducting the analysis. One being that relatively older BMWs with low milage still had sales classifications. This may suggest that brand loyalty and rarity may influence price more that age alone. Comparing this to other brands may help determine if this is due to brand loyalty or long term car quality.

1. **Did you detect any multicollinearity between numeric features? How did you address it?**

Yes. We saw moderate correlation between features such as Engine Size, Horsepower, and Price. This was analyzed using a correlation matrix and variance inflation factor analysis. For models, like Logistic Regression, that are sensitive to multicollinearity, highly correlated features were either removed or combined.

1. **Why did you choose Logistic Regression, Random Forest, and XGBoost for your models?**

This mix of models allowed for both interpretability and strong predictive performance, but each one was chosen because it helped the analysis in a different way.

* Logistic Regression created a simple, interpretable baseline model to look at before diving further into the analysis.
* Random Forest model helped capture the non-linear relationships and interactions without heavy feature engineering.
* XGBoost created a high-performance gradient boosting algorithm that often excels in structured data tasks.

1. **What techniques did you consider (or apply) to address class imbalance?**

Class distributions were considered and reviewed during this process. Techniques like SMOTE and class weighting were considered. Class weights were applied in Logistic Regression and XGBoost to ensure the model paid equal attention to both classes.

1. **How do we interpret the feature importance values from Random Forest or XGBoost?**

Feature importance scores represent how much each feature contributes to reducing classification error. Higher scores indicate stronger predictive influence.

1. **How can dealerships or private sellers use these predictions in real-world sales strategies?**

Predictions like these can help sellers set competitive prices, prioritize marketing for cars with high sales potential, and identify features that could boost resale value. Dealerships could also use this model as an interactive sales prediction tool for real time use.

1. **How would you improve or expand this project if you had more time or data?**

Future improvements could include adding external dataset as mentioned above, testing additional algorithms like Light GBM, and adding data about additional car brands to compare across the industry,