Prediction Model for MMR values of auctioned cars

Developed by: Shantanu Singh

Github:

Linkedin:

Table of Contents

1. Project Background & Objectives
2. Data Overview
3. Exploratory Data Analysis
4. Data Pre-Processing
5. Model Development
6. Results & Assessment
7. Next steps for improvements
8. Annexure

Project Background

An open dataset for auctioned cars in U.S. was obtained from Kaggle. This dataset initially contained over 5.8 lakh+ samples. After cleaning, which involved removing entries with missing information or zero values for either the Manheim Market Report (MMR) or selling price, and eliminating odd samples (e.g. age of the car is one year but odometer rating exceeds 900,000 kms), the dataset was refined to approximately **5.3 lakh samples**. The dataset provides comprehensive details for each sample, including:

- **Vehicle Specifications:** Year of purchase, brand, model, body type, transmission type, and color.
- **Condition & Usage:** Condition rating and odometer reading.
- **Location & Transaction Details:** State of auction, selling price, and date of sale/auction.
- **Manheim Market Report (MMR):** A crucial metric widely used in the U.S. for estimating a vehicle's wholesale price.

Problem Statement: Analyzing the dataset revealed that a significant portion of vehicles (2.71 lakh out of 5.3 lakh) were sold at auction for price lower than MMR, resulting in loss for the dealer.

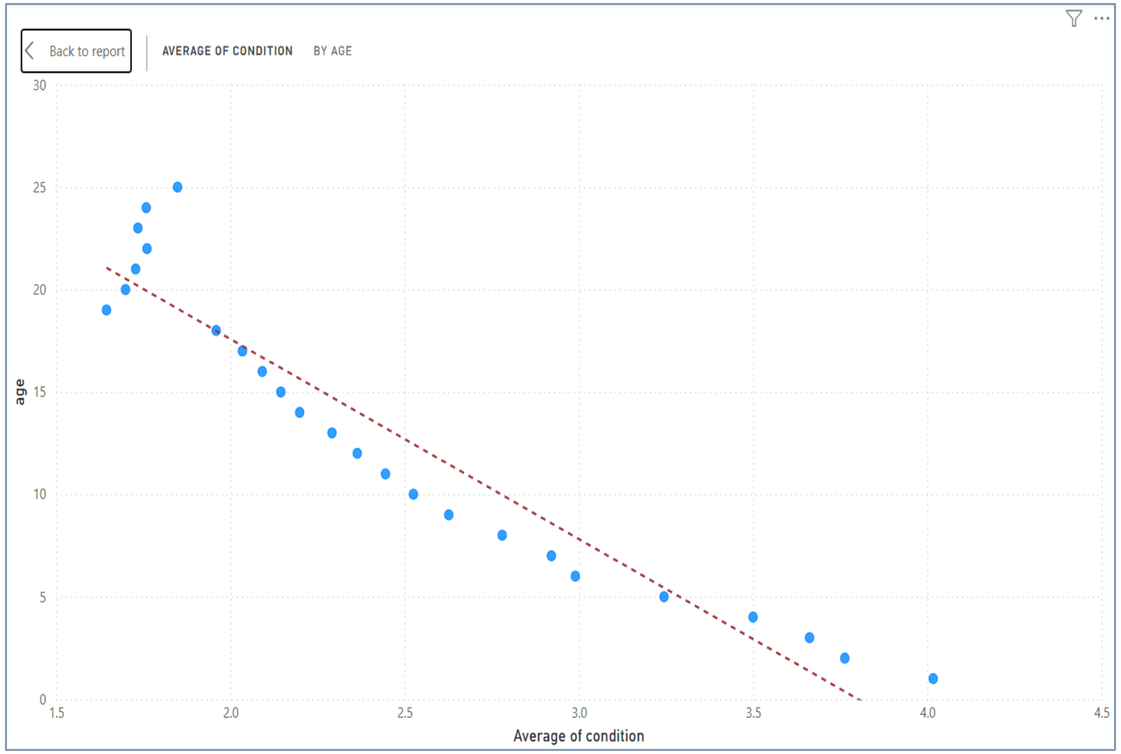
Project Objectives:

1. **Determine key factors** (e.g. vehicle specifications, condition, usage etc.) which influence vehicle’s wholesale (MMR) value
2. **Develop a predictive model** to forecast more accurate MMR value, thereby minimizing probability of loss on the sale of vehicle

Exploratory Data Analysis (EDA)

To reduce dataset complexity, following feature engineering steps were taken:

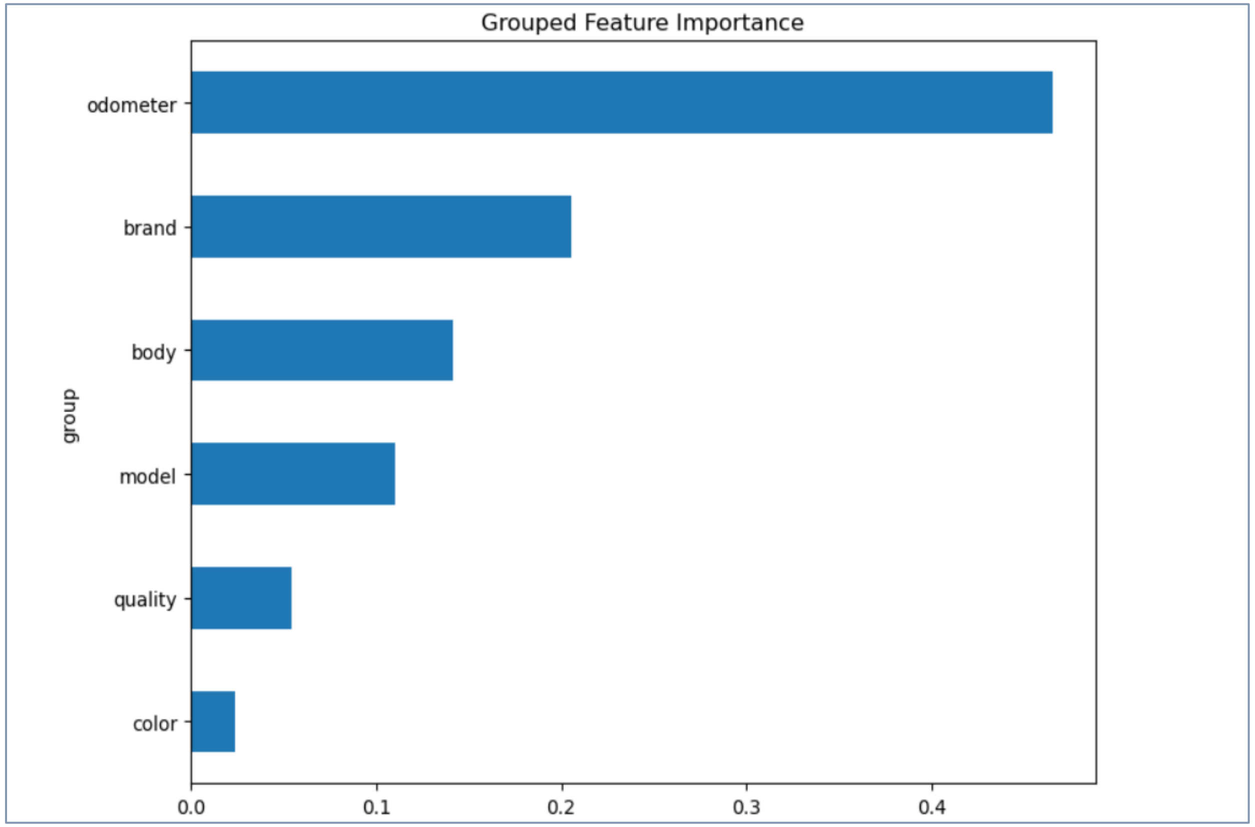
- Car models with similar makes and price points were merged into single sub-categories.
- Body types with very low sample counts were consolidated into 'Others' sub-category.
- Age was derived by subtracting year of purchase from year of sale.
- Final features and number of sub-categories are as follows: Brand (53), Model (181), Body (44), State(39), Condition(Rating: 1 to 5), Odometer(continuous numerical series), Color(19), Age(1 to 25).
- During EDA it was observed that there is a consistent inverse correlation between Age and Condition **(scatter plot below)**. Subsequently, both 'Condition' and 'Age' were scaled using MinMax Scaler in Python and combined into a single feature **'Quality'**.



Note: Additionally, **Principal Component Analysis (PCA)** was performed to assess the correlation between 'Odometer', 'Condition', and 'Age'. Gradient plots revealed that 'Odometer' exhibited a neutral alignment with both principal components, whereas 'Condition' and 'Age' showed some alignment. This suggests that 'Odometer' is not strongly correlated with 'Condition' or 'Age', therefore quality and odometer were not merged together.

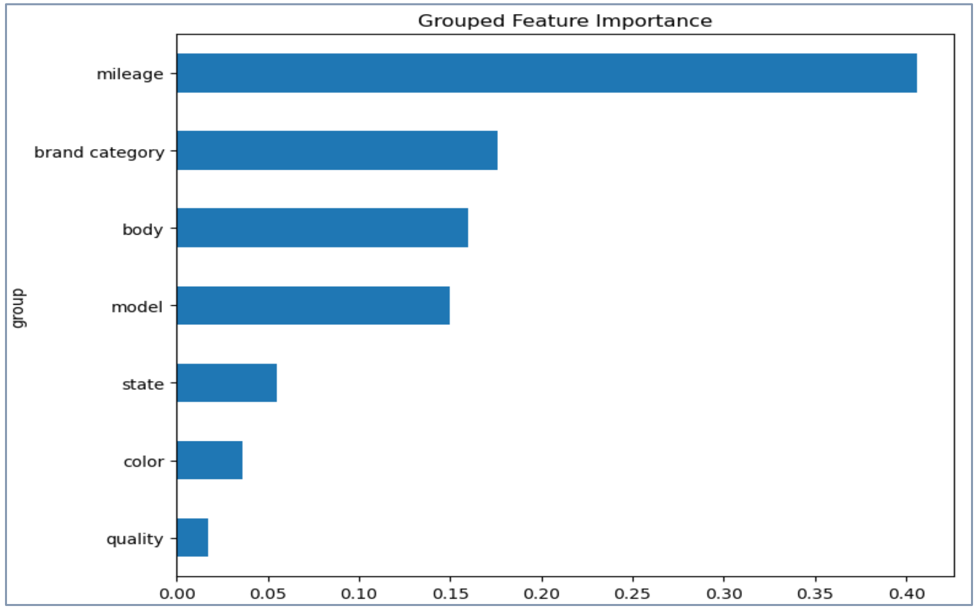
Exploratory Data Analysis (EDA)

- Random Forest Regressor was used in Python to determine most influential features in generating MMR.
- Following encoding/scaling techniques were used for respective features: frequency encoding- *Model*, one hot encoding- *Brand, Body, Color*, minmax scaler- *Odometer, Quality*.
- Random Forest Regressor model was trained using features as input and MMR as the dependent variable. The model utilized 100 estimators.
- Subsequently, feature_importance_ attribute of scikit-learn library was used to extract relative feature importances (depicted in the graph below).
- Odometer was the most significant contributor to MMR, followed by Brand, Body, Model, Quality and finally Color.



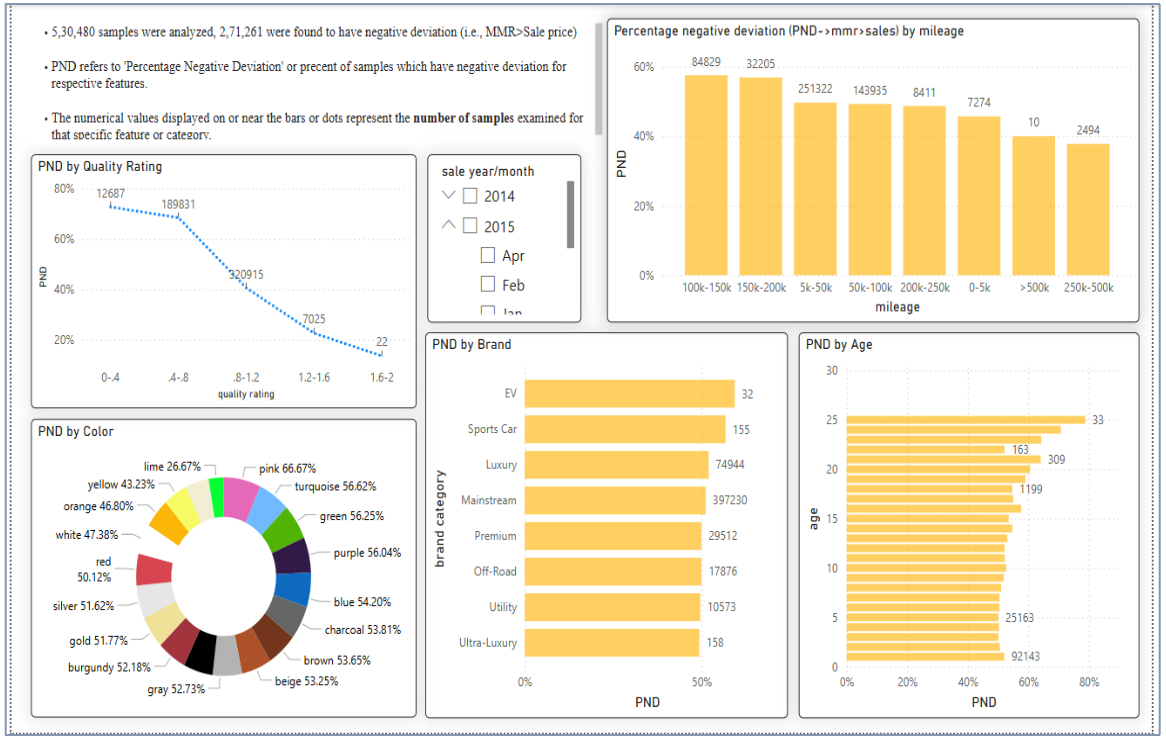
Exploratory Data Analysis (EDA)

- To ensure the model is not biased towards odometer feature, a second round of feature importance analysis was conducted. Since odometer is a continuous feature which can cause it to get selected more frequently for splits than other features, the odometer ratings were encoded into 8 categories, i.e., '0-5k': 1, '5k-50k': 2, '50k-100k': 3, '100k-150k': 4, '150k-200k': 5, '200k-250k': 6, '250k-500k': 7, '>500k': 8 and the feature was renamed to **'Mileage'.**
- Furthermore, the quality feature was encoded into 5 categories: '0-.4': 1, '.4-.8': 2, '.8-1.2': 3, '1.2-1.6': 4, '1.6-2': 5.
- Brands were broadly classified into 8 categories: Mainstream, Premium, Luxury, Ultra-Luxury, Sports, Utility, Off-road and EV.
- **'State'** was included as a new feature in the analysis. Both brand category and state were one hot encoded.
- Random Forest Regressor model was re-trained using 7 features as input and MMR as the dependent variable.
- Mileage (odometer) again emerged as the most significant contributor to MMR, followed by other features, reaffirming the model's estimation in previous run. A notable benefit of these additional encodings was a significant improvement in the model's training time during the second run.
- Weight distribution for each feature importance is as follows: mileage- 41%, brand category-17%, body- 16%, model-15%, state-5%, color-4%, quality-2%.

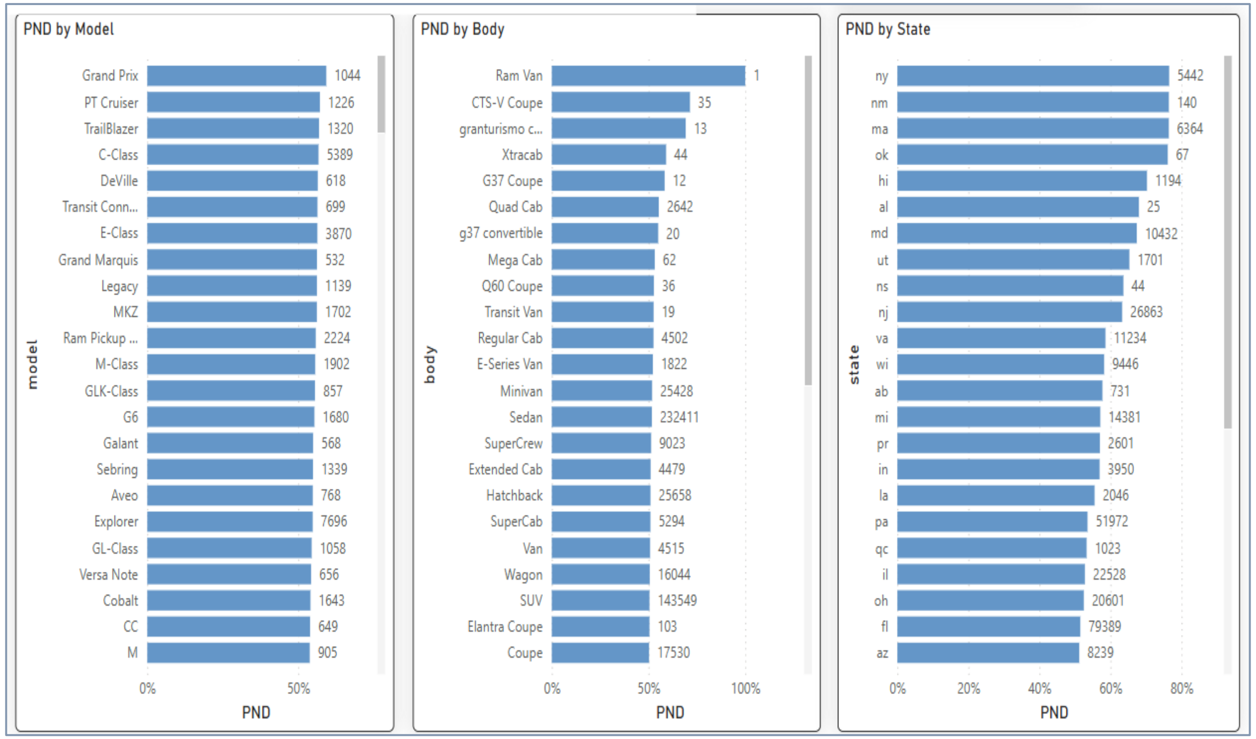


Exploratory Data Analysis (EDA)

- To understand how strongly feature categories are associated with potential losses, **Percentage Negative Deviation (PND)** were calculated for each sub-category across all features by visualizing them in Power BI (as illustrated below).
- PND represents the proportion of samples where the MMR value exceeded the actual sale price (indicating a negative deviation or potential loss). For example, if 39,000 out of 75,000 'Luxury' brand vehicles experienced a negative deviation, the PND for the 'Luxury' category would be 52%.
- These PND values were used in defining a 'risk score' for each category, a concept elaborated in the subsequent data preprocessing section.



- 5,30,480 samples were analyzed, 2,71,261 were found to have negative deviation (i.e., MMR>Sale price)
- PND refers to 'Percentage Negative Deviation' or precent of samples which have negative deviation for respective features.
- The numerical values displayed on or near the bars or dots represent the **number of samples** examined for that specific feature or category.



Data Pre-Processing

- To quantify the risk associated with each category across all seven features, a risk score was computed by combining Percentage Negative Deviation (PND) and sample count.
- The calculation involved two steps:

1. Calculate adjusted risk score:
   - The raw sample count is first normalized by dividing it by 3,00,000.
   - This normalized value and category’s PND value is then summed to together to yield and adjusted score.

Example: The PND value for Premium category in brand feature is 49.96% (.4996 in decimal), sample count is 29,512. Normalized count = 29,512/3,00,000= 0.10 (approx). Adjusted risk score = 0.49 + 0.10 = 0.59 (rounded to 0.60)

2. Find normalized value for risk scores:

All Adjusted Scores for categories of a given feature were then scaled using Min-Max normalization. This transformation ensures the final score falls between 0 and 1, with the highest Adjusted Score mapping to 1. Final normalized value = (adjusted score – min adj score) / (max adj score – min adj score).

Example: Minimum adjusted risk score for brand is 0.50 (Ultra luxury category), and the maximum is 1.84 (Mainstream category). Final normalized value for Premium category is = (0.60-0.50)/(1.82-0.50) = 0.08.

- Final normalized values for risk scores were calculated for all categories across all features. For reference, details of normalized risk scores for some of the features are provided in **annexure**.
- The normalized risk scores for each category were appended to their corresponding rows/samples using VLOOKUP function. A Composite **Risk Score** was then generated for each individual vehicle by multiplying each category's risk score by its respective feature's weight (as determined in Slide 6, e.g., 41% for Mileage, 17% for Brand) and summing these weighted contributions.

Data Pre-Processing

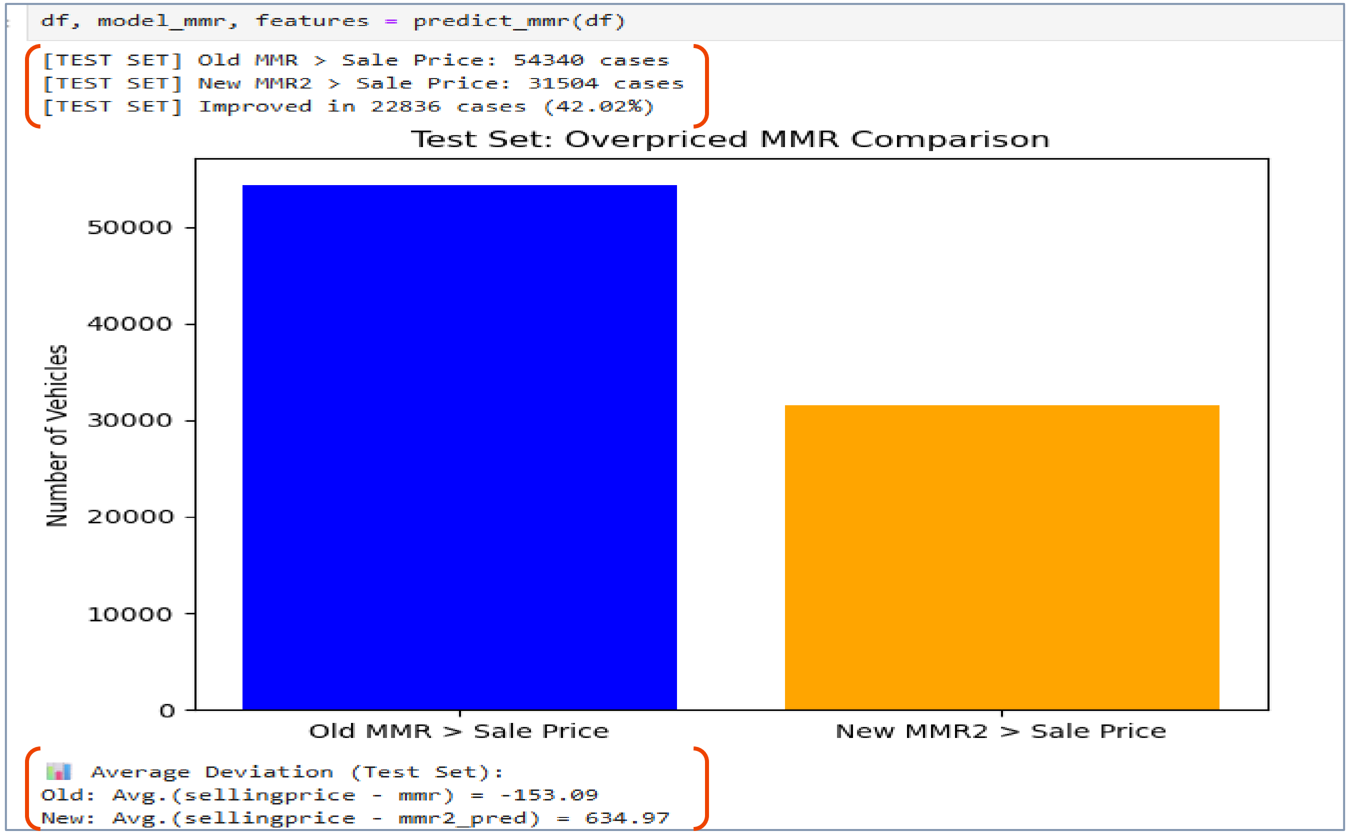
- This Composite Risk Score was instrumental in deriving a more optimized MMR value (termed **'MMR 2'**) designed to mitigate potential losses on vehicle sales without excessively reducing the estimated wholesale price. The adjustment was applied based on the following logic:
  - If the Composite Risk Score for a vehicle was greater than 0.9, 'MMR 2' was set to 91% of its existing MMR.
  - If the Composite Risk Score was between 0.8 and 0.9, 'MMR 2' was set to 92% of its existing MMR. And so forth for the remaining risk score ranges.

Model Development

- A predictive model was developed to forecast a more accurate MMR value, referred to as **MMR 2**. The dataset was prepared with the following feature engineering and encoding techniques:
  - Model: Frequency encoding
  - Brand category, body, color, state: One hot encoding
  - Mileage, Quality: Binning
  - Risk Score and Original MMR: These continuous numerical features were used without encoding, as Risk score was already scaled between 0 and 1, and MMR served as a foundational variable for the target.
- The dataset was split into an 80:20 ratio for training and testing, respectively. **XGB Regressor** was selected for model training and prediction, leveraging its performance on structured data.
- The model's performance was assessed by measuring its impact on profitability and predictive accuracy. Key metrics included:
  - **Reduction in Negative Deviation:** The change in the number of cases where the predicted MMR 2 exceeded the actual selling price, indicating a reduction in potential losses.
  - **Average Deviation:** The mean difference between the selling price and the predicted value, evaluated before and after model application.
  - **Accuracy Metrics:** Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared.
- Tools and libraries used during EDA and model development:
  - Software: Python, Power BI, Excel
  - Libraries: xgboost, sklearn, pandas, numpy, matplotlib, shap
  - Regression models: XGBRegressor, Random Forest, Linear Regression
  - Scalers: Minmax scaler, Standard scaler

Results & Assessment

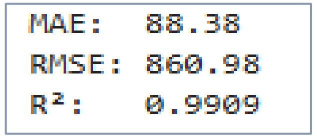
- The predictive model was tested on 20% of the dataset, comprising **106,096 samples**. The results demonstrate a significant improvement in profitability and prediction accuracy. A substantial **42.02% reduction** was achieved in cases where the predicted MMR value exceeded the sale price, mitigating potential losses.
- The average deviation between selling price and MMR saw a remarkable shift from -$153.09 (using Original MMR) to **$634.97** (using Predicted MMR 2) within the tested samples, indicating a substantial improvement in average profit margin.
- MAE of **88.38** indicates a minimal deviation in predicted values, especially given the large-scale values of MMR.
- While the RMSE of 860.98 is notably higher than MAE, which indicates there are some extreme outlier cases in the prediction.
- An R-squared value of **0.9909** confirms that the model is highly accurate at predicting the target value, based on the provided input features.



```
df, model_mmr, features = predict_mmr(df)
```

```
[TEST SET] Old MMR > Sale Price: 54340 cases
[TEST SET] New MMR2 > Sale Price: 31504 cases
[TEST SET] Improved in 22836 cases (42.02%)
```

```
📊 Average Deviation (Test Set):
Old: Avg.(sellingprice - mmr) = -153.09
New: Avg.(sellingprice - mmr2_pred) = 634.97
```



```
MAE:   88.38
RMSE:  860.98
R²:    0.9909
```

Next steps for improving the predictive model

Following strategies can be used to further improve the predictive model:

- **Integrate Macroeconomic and Market Dynamics:** Incorporate features representing demand/supply trends and broader economic indicators specific to the time of sale. This will provide the model with crucial context on market liquidity and sentiment, thereby refining the risk scoring.
- **Outlier Analysis and Robustness:** Conduct further analysis to identify and understand the characteristics of outlier cases that contribute to extreme deviations in predictions.
- **Dealer-Specific Information:** Introduce 'Dealer Name' as a feature in the dataset. Dealers are a significant factor influencing sale prices, and including this information will allow the model to capture dealer-specific effects, leading to improved risk score and a more accurate target MMR.

Thank You

Annexure- Normalized Risk Scores

mileage	PND	Count of samples	adj pnd mileage	adj mileage norm
5k-50k	49.66%	251322	0.67	1.00
50k-100k	49.28%	143935	0.49	0.62
100k-150k	57.56%	84829	0.43	0.50
150k-200k	56.90%	32205	0.34	0.31
200k-250k	48.70%	8411	0.26	0.14
0-5k	45.71%	7274	0.24	0.10
250k-500k	37.81%	2494	0.19	0.00
>500k	40.00%	10	0.20	0.01

brand category	PND	Count of samples	adj pnd brand	adj brand norm
Mainstream	51.15%	397230	0.92	1.00
Luxury	52.00%	74944	0.38	0.21
Premium	49.96%	29512	0.30	0.08
Off-Road	49.95%	17876	0.28	0.05
Utility	49.63%	10573	0.27	0.03
Ultra-Luxury	49.37%	158	0.25	0.00
Sports Car	56.77%	155	0.28	0.06
EV	59.38%	32	0.30	0.07

quality rating	PND	Count of samples	adj pnd quality	adj count norm
.8-1.2	40.66%	320915	0.74	1.00
.4-.8	68.46%	189831	0.66	0.88
0-.4	72.70%	12687	0.38	0.47
1.2-1.6	22.68%	7025	0.13	0.08
1.6-2	13.64%	22	0.07	0.00