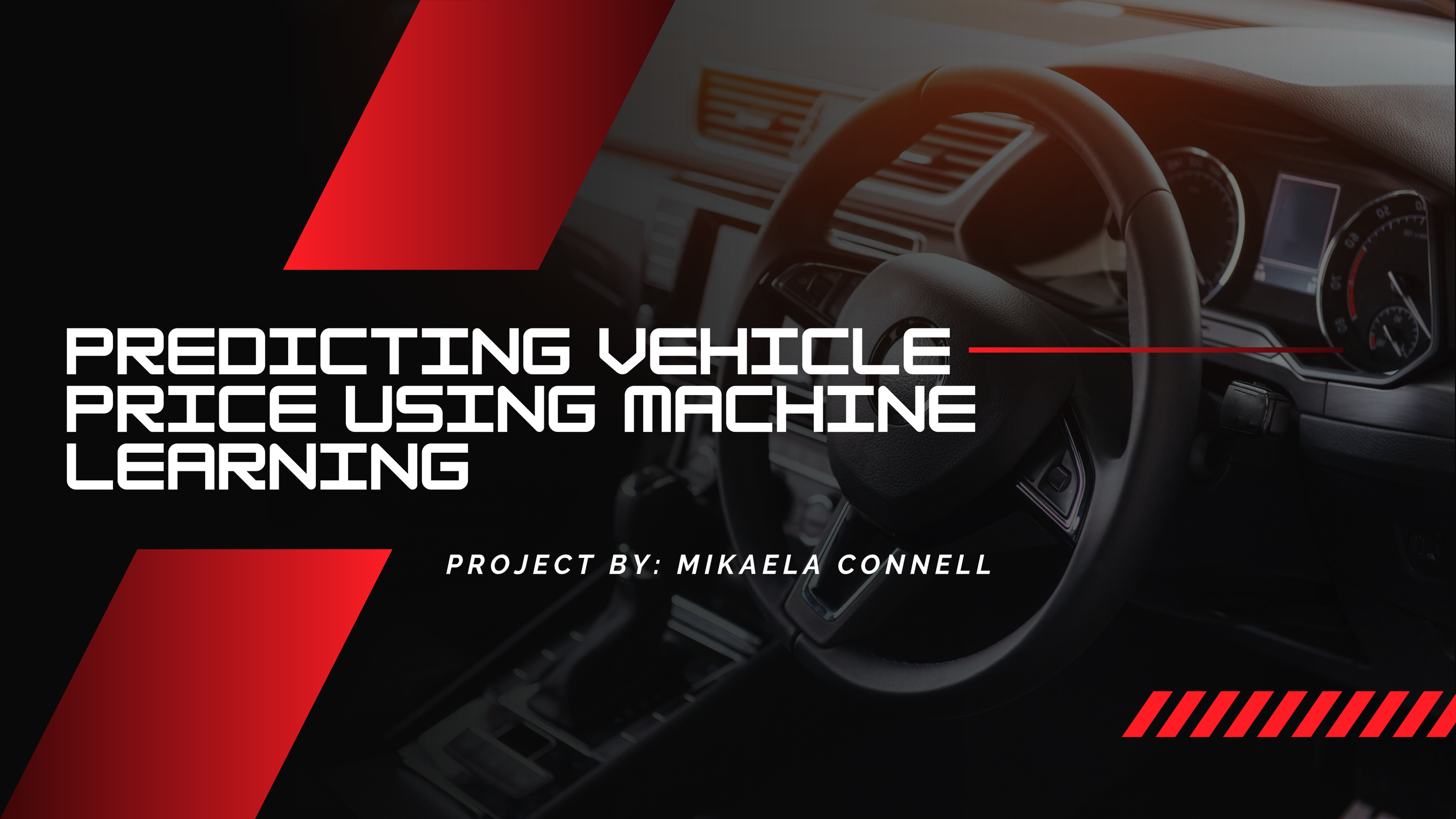


PREDICTING VEHICLE PRICE USING MACHINE LEARNING

PROJECT BY: MIKAELA CONNELL

PROJECT OVERVIEW

Business Goal:

- Develop a predictive model that accurately estimates vehicle prices based on key attributes
- Help car dealerships, online marketplaces, or individual sellers to set competitive and profitable prices based on data-driven insights

Data Analysis Performed:

- Analyzed data from over 550,000 vehicles
- Key features:
  - Manufacturing year, condition, odometer, body type, transmission, condition, & mileage
  - transaction data: selling price, sale date

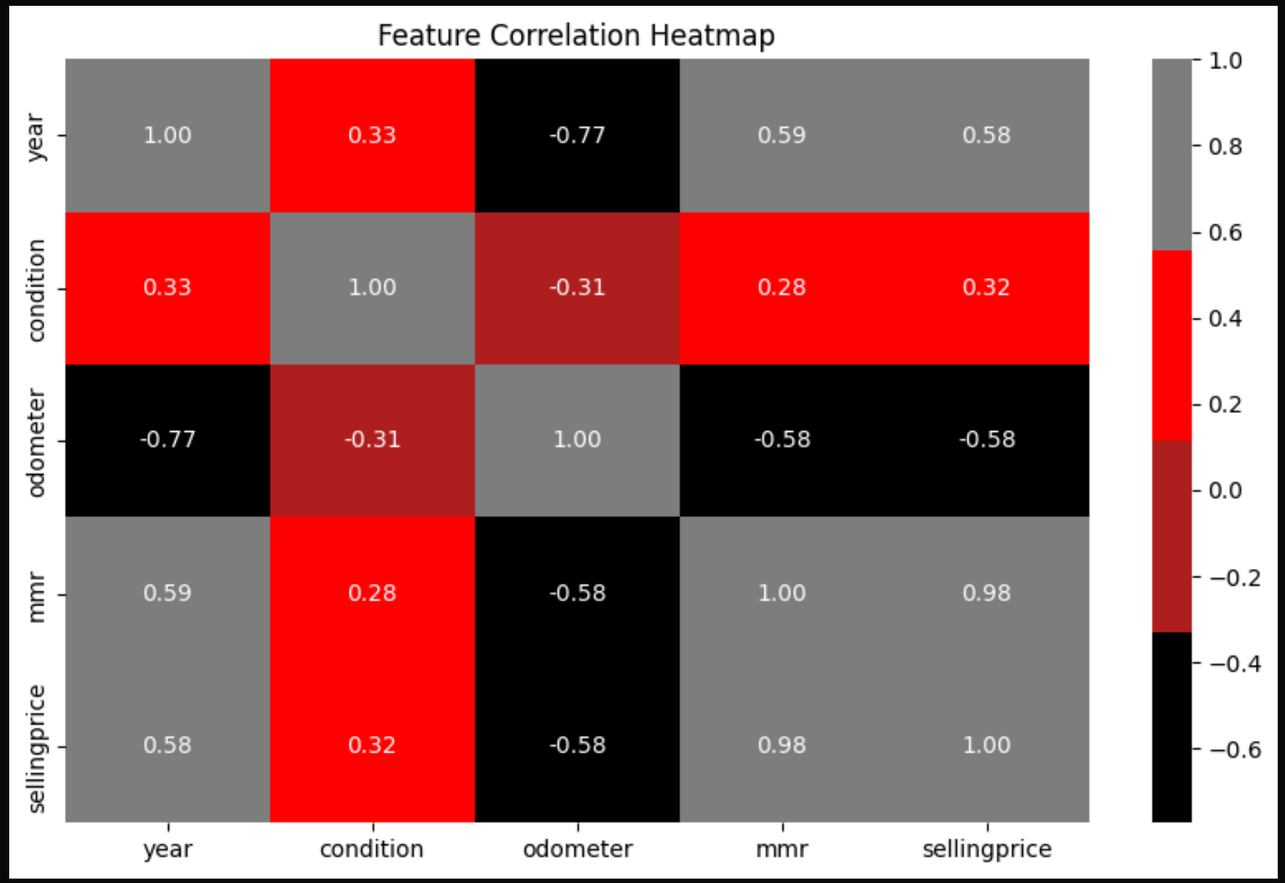
CORRELATION ANALYSIS

Positive Correlation:

- **MMR & Selling Price:** Manheim Market Report price is a key indicator of wholesale car values
- **Manufacturing Year & Selling Price:** Newer cars tend to have higher selling prices
- **Condition Rating & Selling Price:** Well-maintained cars sell for higher prices

Negative Correlation:

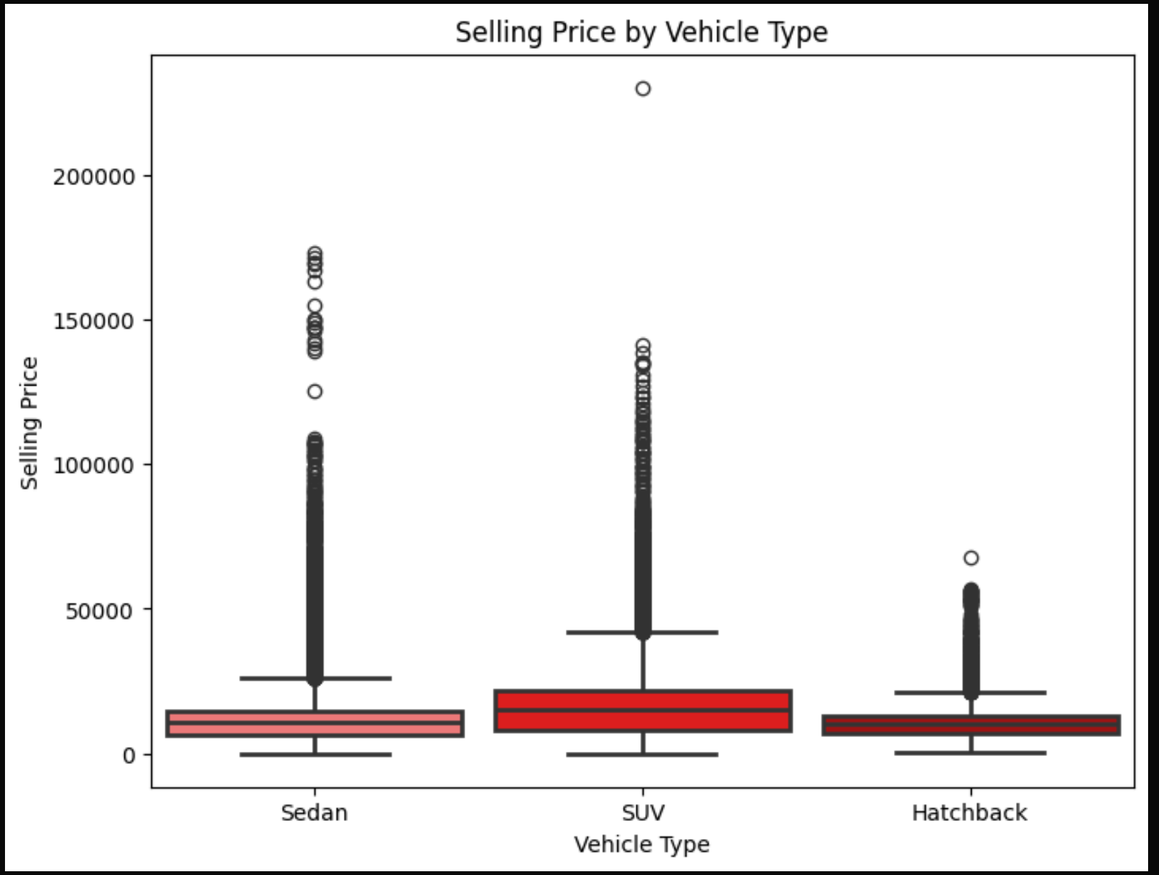
- **Odometer reading & Selling Price:** Higher mileage reduces vehicle value



SUMMARY OF STATISTICAL ANALYSIS

Comparative Analysis (T-test & ANOVA)

- T-test showed that **automatic transmission vehicles tend to have higher selling prices than manual vehicles**
- ANOVA analysis indicated significant price difference across body types, with **SUVs and Sedans generally priced higher than hatchbacks**



CLASSICAL MACHINE LEARNING APPROACH

Linear Regression

- R^2 = 0.9679
- MAE = $1,082

Random Forest

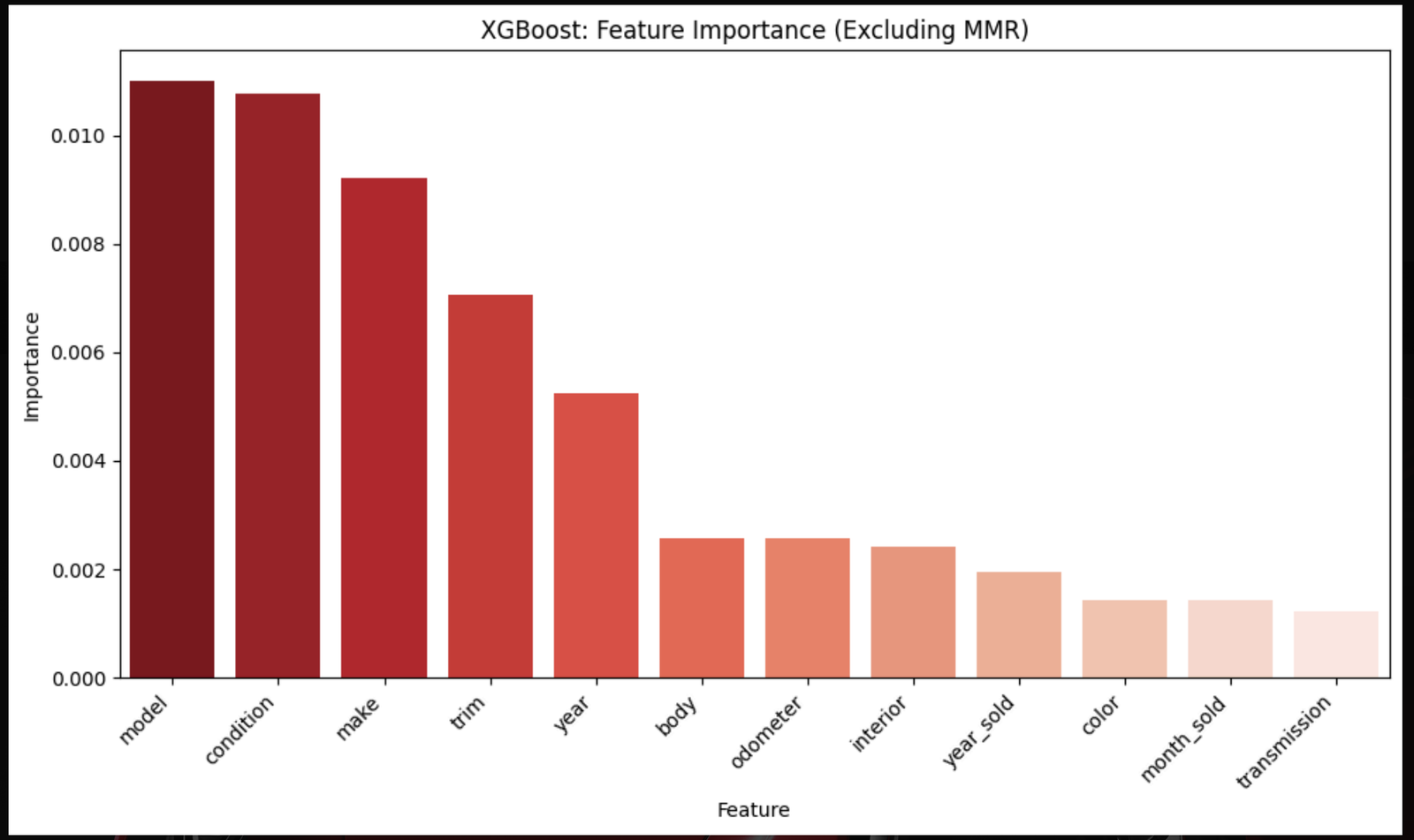
- R^2 = 0.9706
- MAE = $1,056
- Performs slightly better than linear regression

XGBoost

- R^2 = 0.9670
- MAE = $1,037
- Best performing model

- XGBoost had the lowest MAE, indicating the smallest prediction errors
- The XGBoost R^2 score was slightly lower than Random Forest but still showed strong predictive performance

GRADIENT BOOSTING MODEL (XGBOOST): FEATURE IMPORTANCE



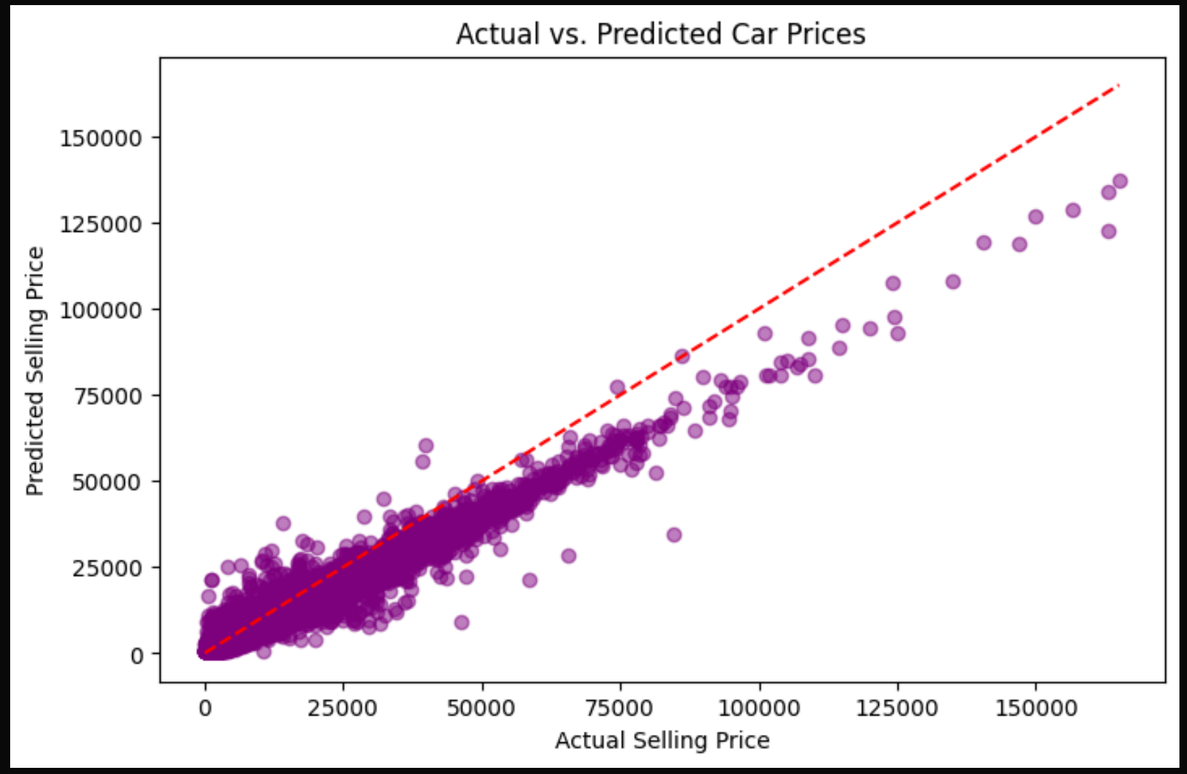
DEEP LEARNING APPROACH: ARTIFICIAL NEURAL NETWORK (ANN)

Network Architecture

- **Custom-built ANN**
- **Input layer:** all numerical features
- **Hidden layers**: multiple dense layers with ReLU activation, dropout layers to prevent overfitting, & batch normalization to stabilize training
- **Output layer**: a single neuron with linear activation for price prediction

Model Performance

- **Test Mean Average Error** (MAE): $2,606, a higher error value compared to XGBoost & Random Forest Models
- **Challenges identified**: Some overfitting was observerd (Validation Loss increased after a few epochs), indicating need for further regularization & hyperparameter refinement



SUMMARY & FUTURE CONSIDERATIONS

Potential applications of best performing models:

- Car resale pricing tools
- Price recommendations for dealerships
- Auction price forecasting

Future improvements:

- More detailed feature engineering to improve accuracy
- Incorporate external factors such as demand trends

REFERENCES

https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data/data

https://www.nasdaq.com/articles/car-price-predictions-2024:-proof-the-auto-market-crash-has-begun