Final Report

Problem Statement: Accurately predicting house prices is crucial in the real estate market, but it remains challenging due to various factors. The goal of this project is to develop a machine learning model that can reliably predict house prices based on relevant features, ultimately providing valuable insights for stakeholders in the real estate market.

Summary of Variables and Dataset Size

Dataset Overview:

- **Number of Rows:** 1,460
- **Number of Columns:** 81
- **Target Variable:** 'SalePrice'

Variables:

Variable Name	Description
Id	Identification number
MSSubClass	The building class
MSZoning	The general zoning classification
LotFrontage	Lot frontage in the feet
LotArea	Lot size in square feet
Street	Type of road access
Alley	Type of alley access
LotShape	The general shape of the property
LandContour	The flatness of the property

Utilities	Type of utilities available
Neighborhood	Physical locations within Ames city limits
HouseStyle	The style of the house
OverallQual	Overall material and finish quality
OverallCond	Overall condition rating
YearBuilt	Original construction date
YearRemodAdd	Remodel date
RoofStyle	Type of roof
RoofMatl	Roof material
Exterior1st	Exterior covering on the house
Exterior2nd	Exterior covering on the house (if more than one)
MasVnrType	Masonry veneer type
MasVnrArea	Masonry veneer area in square feet
ExterQual	Exterior material quality
ExterCond	Present condition of the material on the exterior
Foundation	Type of foundation
BsmtQual	Height of the basement
BsmtCond	The general condition of the basement

BsmtExposure	Walkout or garden-level basement walls
BsmtFinType1	Quality of basement finished area
BsmtFinSF1	Type 1 finished square feet
TotalBsmtSF	Total square feet of basement area
Heating	Type of heating
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system
1stFlrSF	First-floor square feet
2ndFlrSF	Second-floor square feet
GrLivArea	Above grade (ground) living area square feet
BsmtFullBath	Basement full bathrooms
FullBath	Full bathrooms above grade
BedroomAbvGr	Bedrooms above grade
KitchenQual	Kitchen Quality
TotRmsAbvGrd	Total rooms above grade (excluding bathrooms)
GarageType	Garage location
GarageYrBlt	Year garage was built

GarageFinish	The interior finish of the garage
GarageCars	Size of garage in car capacity
GarageArea	Size of garage in square feet
PoolArea	Pool area in square feet
MiscFeature	Miscellaneous features not covered in other categories
SaleType	Type of sale
SaleCondition	Condition of sale
SalePrice	The property's sale price in dollars

For the full list of variables, refer to the Kaggle dataset page.

Approach:

1. **Data Wrangling:** The dataset was thoroughly cleaned, missing values were addressed, and significant outliers were mitigated to prepare the data for analysis.
    - **Handling Missing Values and Outliers**
        - **Missing Values:**

          Strategy: Missing values were handled based on the nature of the variable. For instance, missing values in categorical variables were replaced with a new category 'None' or the most frequent category. For numerical variables, missing values were often filled with the median or mean, or sometimes with zero if the variable logically supported it.

          Percentage: The dataset has many missing values in some columns. For instance, 'PoolQC' has around 99% missing values, 'MiscFeature' about 96%, 'Alley' about 94%, and 'Fence' about 80%.

        - **Outliers:**

Identification: Outliers were identified using visualizations like box plots and statistical methods.

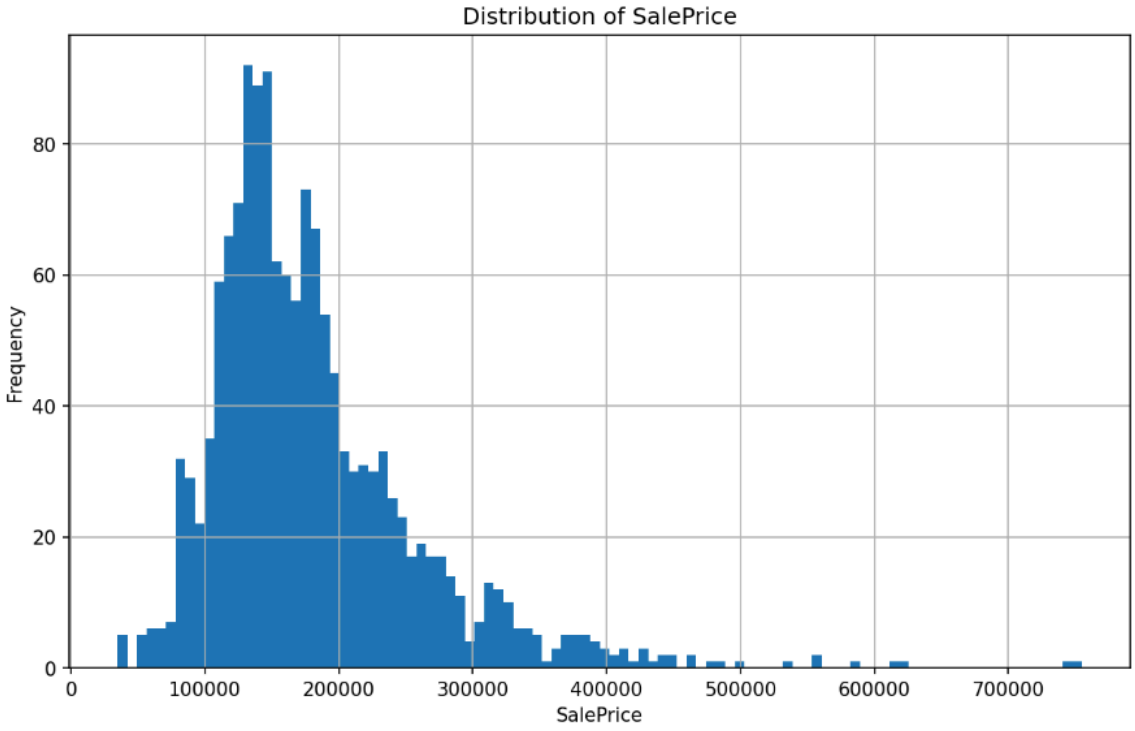
Mitigation: Outliers were handled by capping them at a certain threshold or transforming the data. For example, the 'GrLivArea' variable had some extremely capped outliers.

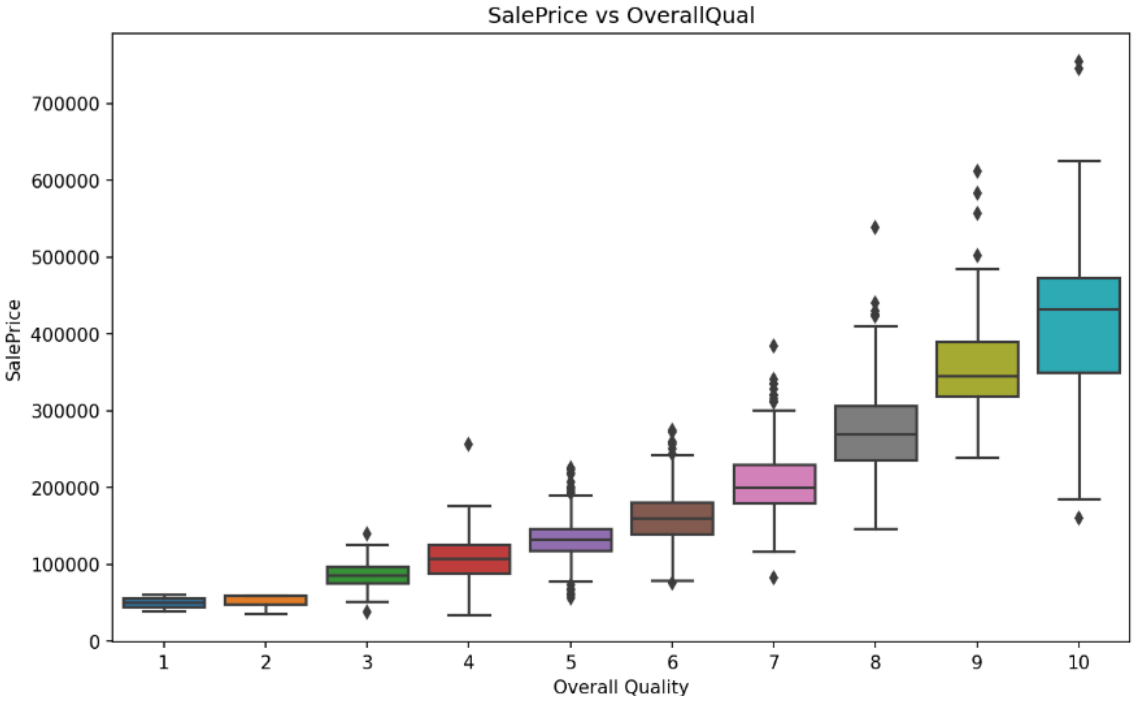
- **Percentage Representation**

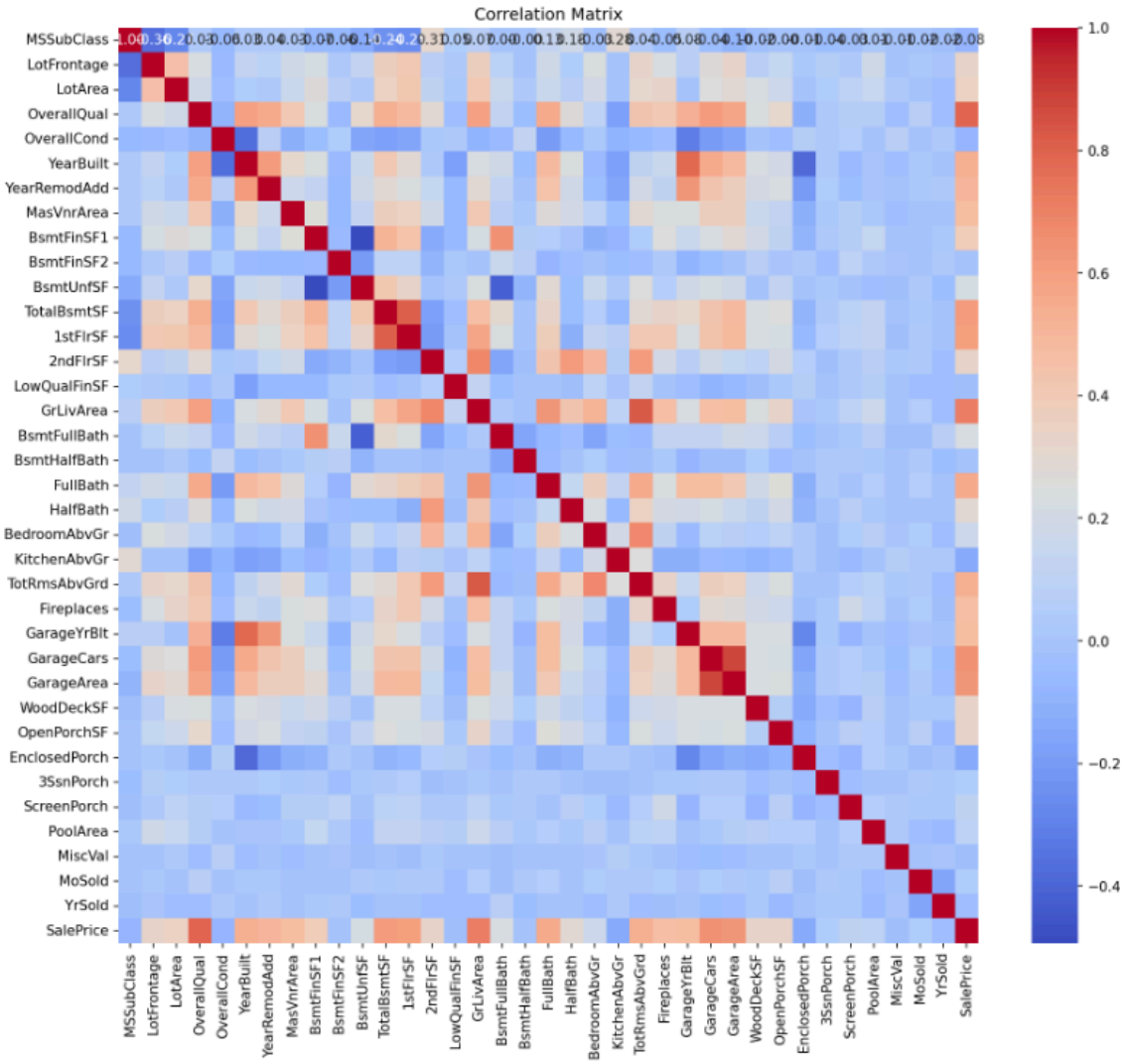
  Missing Values: Some columns with high percentages of missing values were considered for removal. Columns with less than 80% missing data were generally filled with a constant or a meaningful imputed value.

  Outliers: Outliers represented a small percentage but had a significant impact on model performance, thus necessitating their treatment.

2. **Exploratory Data Analysis (EDA):** A comprehensive examination of the dataset was conducted to understand its characteristics and prepare it for modeling endeavors.




SalePrice vs OverallQual
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1
2
3
4
5
6
7
8
9
10
Overall Quality



3. **Feature Engineering:** New variables were created based on existing ones to enrich the dataset, and steps such as rare analysis, label encoding, and scale standardization were performed to enhance the dataset's effectiveness.

   To facilitate the machine learning process, enhance the predictive power of machine learning algorithms, and achieve better results, new features have been created from raw data. When creating new features, the existing features were carefully examined, and in cases where features had strong correlations with each other or where it made logical sense to combine two pieces of data, the two features were merged into a single feature to reduce the number of features.

   For instance,

The **'1stFlrSF'** and **'2ndFlrSF'** features were combined to create a new feature called **'NEW_TotalFlrSF'**. The '1stFlrSF' and '2ndFlrSF' features were then dropped.

The **NEW_HouseAge** feature was created by subtracting **YearBuilt** from **YrSold**.

The **NEW_PorchArea** feature was created by summing up the **OpenPorchSF**, **EnclosedPorch**, **ScreenPorch**, **3SsnPorch**, and **WoodDeckSF** features.

4. **Modeling:** Various machine learning algorithms including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, KNN, SVM, and XGBoost were experimented with. Based on evaluation metrics, XGBoost was identified as the best-performing model.

Findings:

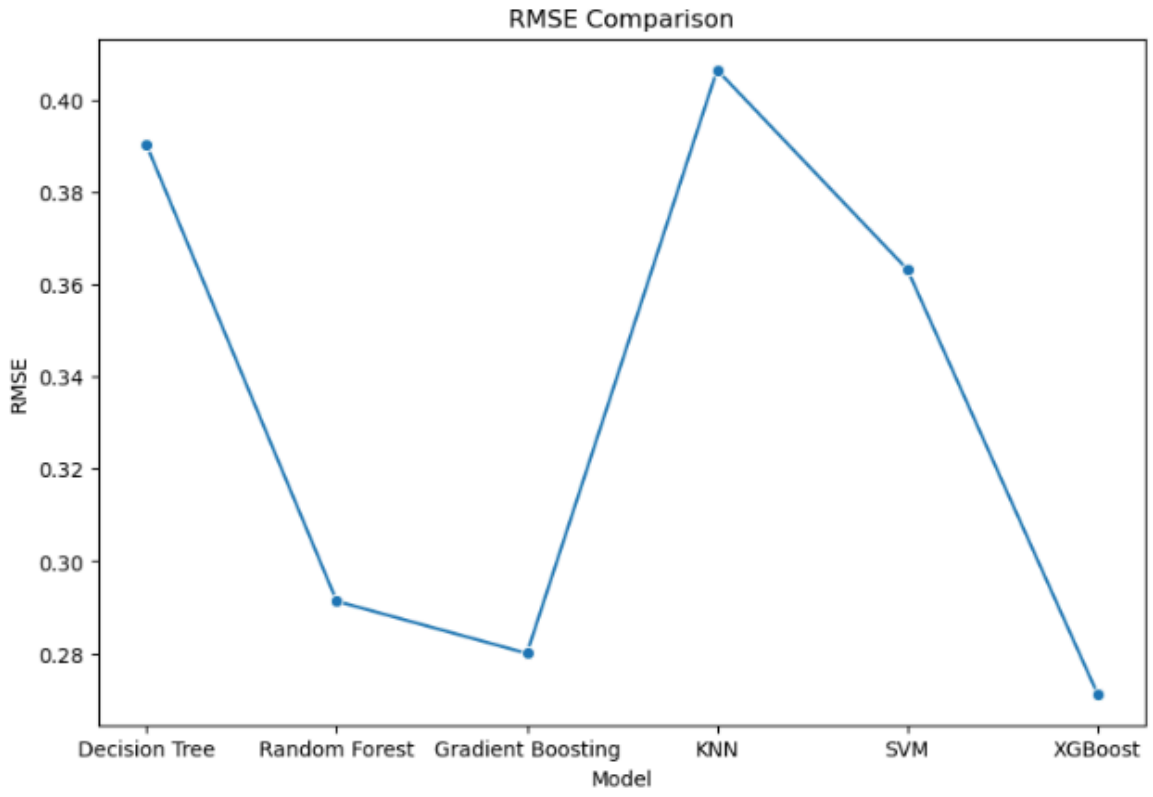
- Linear Regression performed poorly, while Decision Tree, Random Forest, Gradient Boosting, KNN, SVM, and XGBoost showed varying performance degrees.
- XGBoost emerged as the top performer with excellent performance metrics, making it the recommended model for predicting house prices in this scenario.

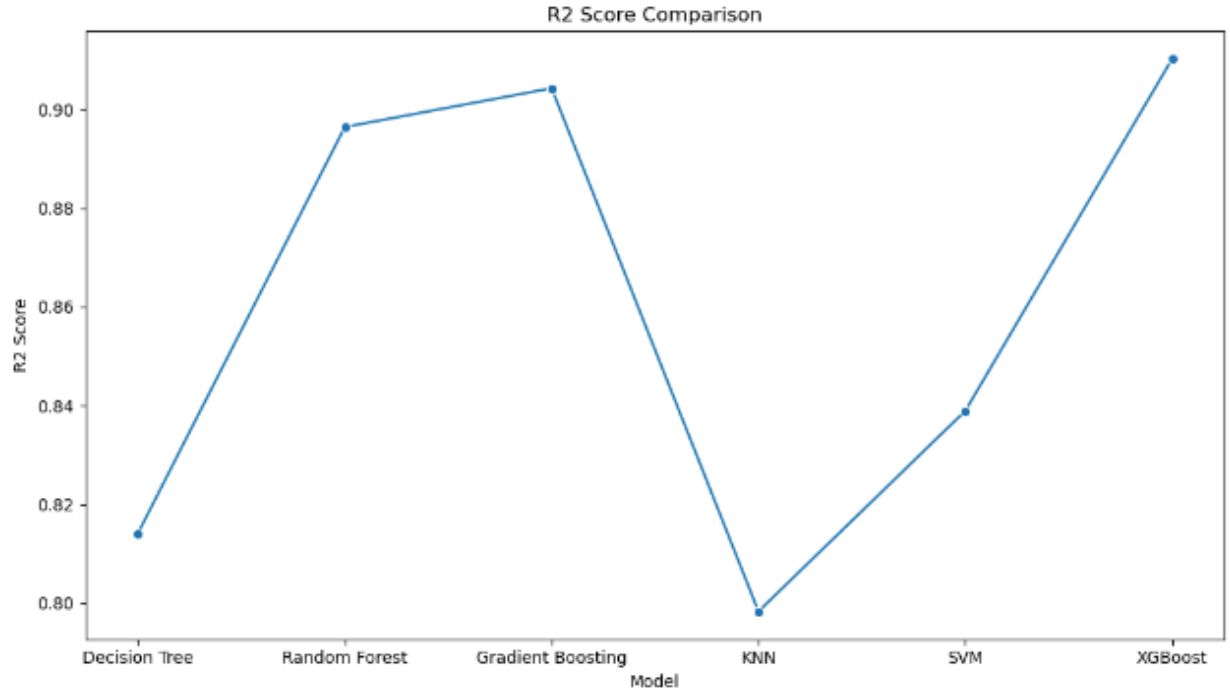
Recommendations:

1. **Utilize XGBoost Model:** Implement the XGBoost model for predicting house prices in real-time scenarios, as it demonstrated superior performance compared to other models.
2. **Continuous Model Monitoring:** Regularly monitor the model's performance and retrain it with updated data to ensure its accuracy and reliability over time.
3. **Further Research:** Explore advanced techniques for feature engineering and model optimization to potentially enhance the model's performance even further.

Model Metrics:

- Model: XGBoost
- Features: Various features related to residential properties
- Parameters: Optimized hyperparameters
- Performance Metrics: Low errors, high R2 score, and excellent performance in predicting house prices.


RMSE Comparison
RMSE
0.40
0.38
0.36
0.34
0.32
0.30
0.28
Decision Tree
Random Forest
Gradient Boosting
KNN
SVM
XGBoost
Model

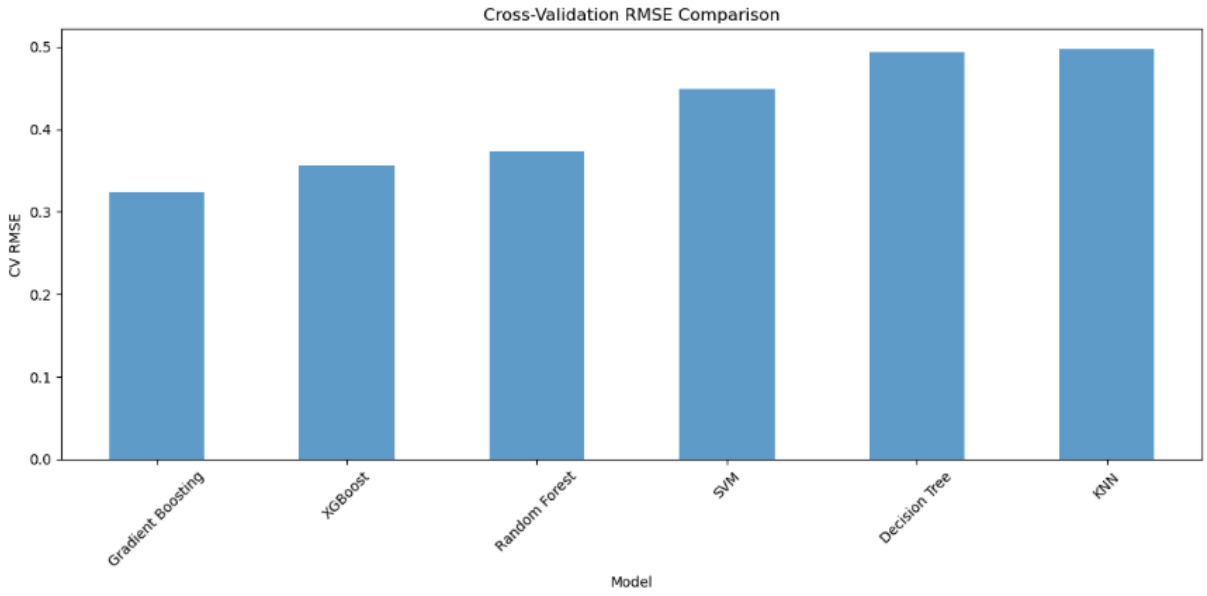

R2 Score Comparison
R2 Score
0.90
0.88
0.86
0.84
0.82
0.80
Decision Tree
Random Forest
Gradient Boosting
KNN
SVM
XGBoost
Model

	CV RMSE	MAE	MSE	RMSE	R2
Decision Tree	0.479816	0.278857	0.152220	0.390154	0.814162
Random Forest	0.377677	0.190229	0.084867	0.291320	0.896390
Gradient Boosting	0.328736	0.182895	0.078422	0.280039	0.904258
KNN	0.496686	0.268435	0.165182	0.406426	0.798337
SVM	0.448375	0.235250	0.131934	0.363228	0.838928
XGBoost	0.355961	0.178966	0.073584	0.271264	0.910165

Explanation of metrics:

- RMSE (Root Mean Squared Error): A lower RMSE indicates a better fit of the model as it measures the average magnitude of the errors between predicted and actual values. XGBoost model has the lowest RMSE value.
- R2 Score (Coefficient of Determination): A higher R2 score indicates a better fit as it represents the proportion of the variance in the dependent variable that is predictable from the independent variables. XGBoost model has the highest R2 score value.

Conclusion: This project successfully addressed the challenge of predicting house prices accurately by developing and evaluating machine learning models. The XGBoost model emerged as the most effective solution, providing valuable insights for stakeholders in the real estate market. Further refinement and continuous model monitoring can ensure its continued relevance and usefulness in real-world applications.



Notebooks and Scripts:

1. Problem Statement
2. Data Wrangling Notebook
3. Exploratory Data Analysis (EDA) Notebook
4. Feature Engineering Notebook
5. Modeling Notebook
6. Model Metrics File