AUTOMATED MACHINE LEARNING ANALYSIS REPORT

Generated: February 15, 2026 at 23:55
Report Type: Full ML Pipeline Analysis

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

This report presents a comprehensive analysis of an automated machine learning pipeline executed on February 15, 2026 at 23:55. The analysis encompassed data exploration, preprocessing, feature engineering, model selection, and performance evaluation.

Key Findings

Metric	Value
Best Model	GradientBoostingRegressor
Model Score	0.9066
Models Evaluated	3

2. DATA EXPLORATORY ANALYSIS

2.1 Dataset Overview

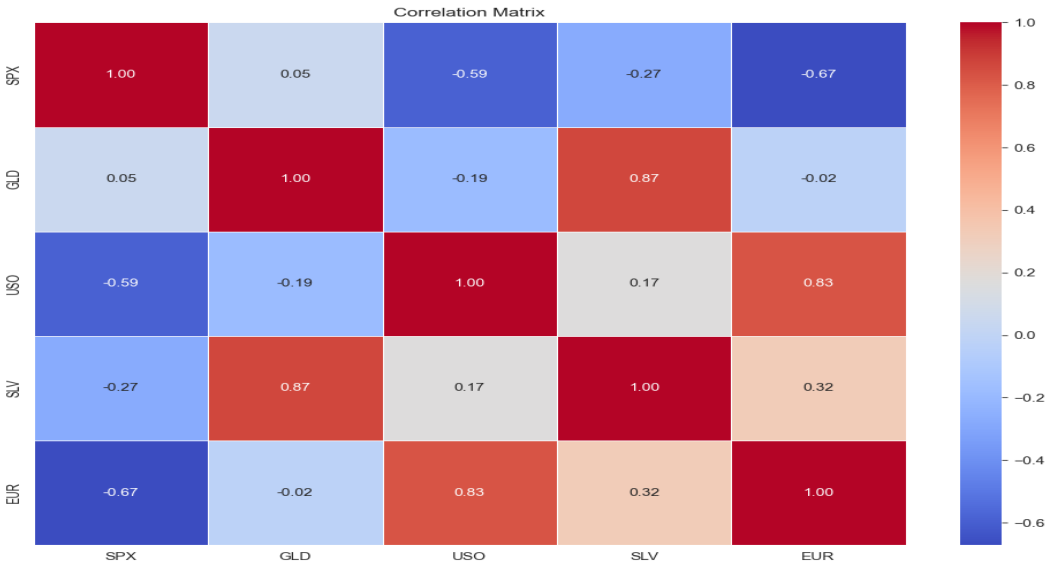
EXECUTIVE SUMMARY This dataset contains information on 8,128 vehicles, including their selling prices, kilometers driven, number of seats, and other attributes. KEY DATA INSIGHTS - The mean selling price is $638,272, with a standard deviation of $806,253, indicating a significant range in prices across the data. This suggests that while some vehicles are sold at very high prices, others are sold at much lower prices. - The minimum and maximum kilometers driven indicate that there is a wide range of driving experiences among vehicle owners. On one hand, some drivers have logged only 1 km, suggesting occasional or short trips, while others have accumulated over 2 million km, indicating extensive use of the vehicles. - The number of seats ranges from 2 to 14, with most vehicles having 5 seats. This could indicate that there is a mix of compact cars and larger vehicles in the dataset. DATA QUALITY & RISKS - There are no missing values for any of the main attributes (year, selling price, km driven, seats), but there are some missing values for the fuel type attribute. - Outliers exist in the year, selling price, km driven, and seats columns. Specifically, one vehicle has been on the road for 81 years, one vehicle was sold at $600, another vehicle has logged over 2 million km, and a third vehicle has all four seats. These outliers could potentially skew results from analysis or modeling. CONCLUSION The dataset provides valuable insights into vehicle sales trends, driving habits, and customer preferences, but it also presents some quality risks due to missing values and outliers that must be carefully addressed in order to draw accurate conclusions about the data subject.

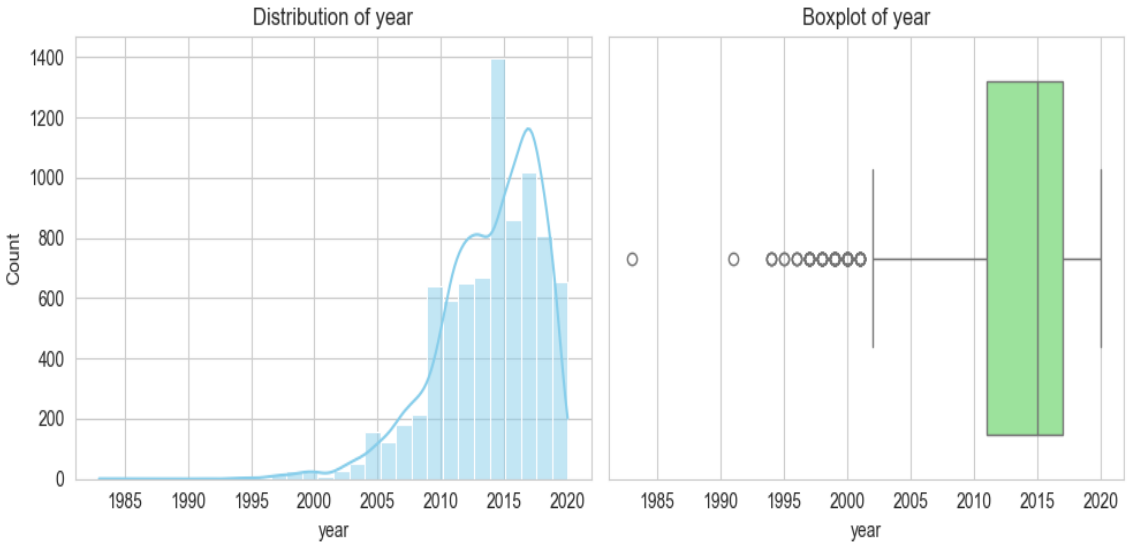
2.2 Data Quality Assessment

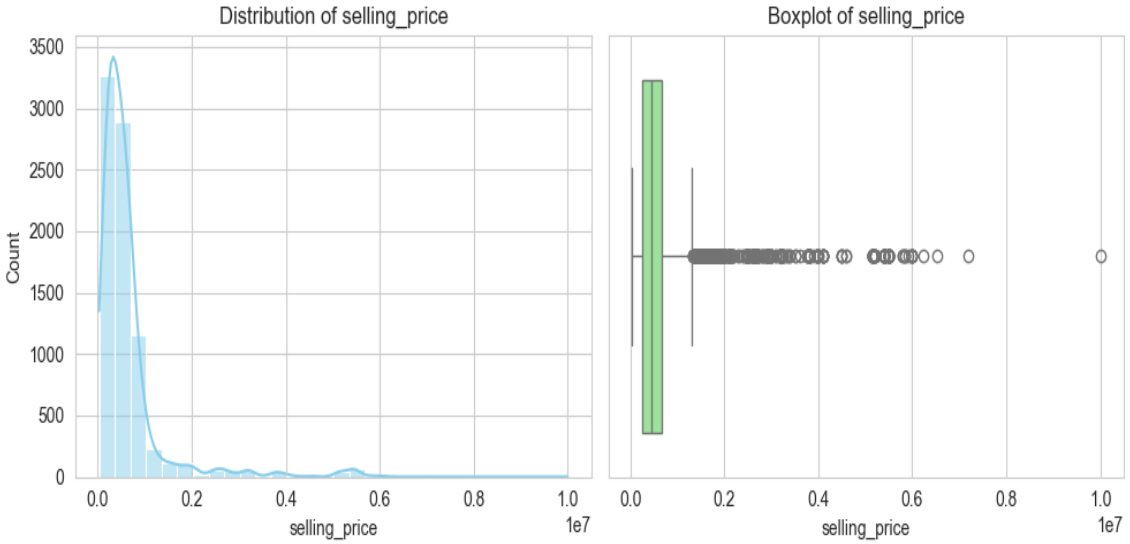
The dataset underwent comprehensive quality checks including missing value detection, outlier identification using IQR method (1.5x threshold), and distribution analysis. All identified issues were documented and addressed in the preprocessing phase.

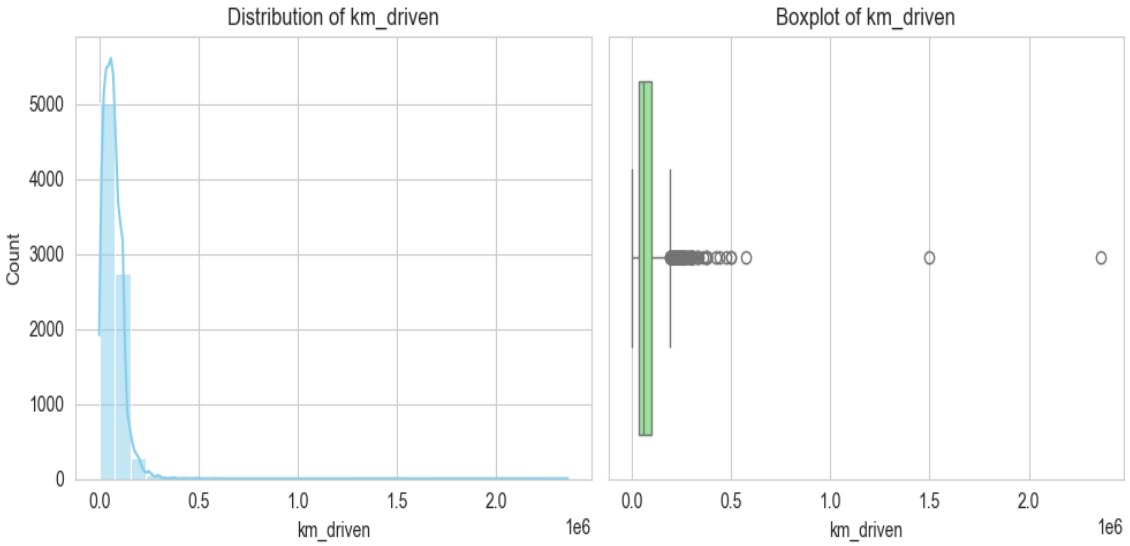
2.3 Key Data Visualizations

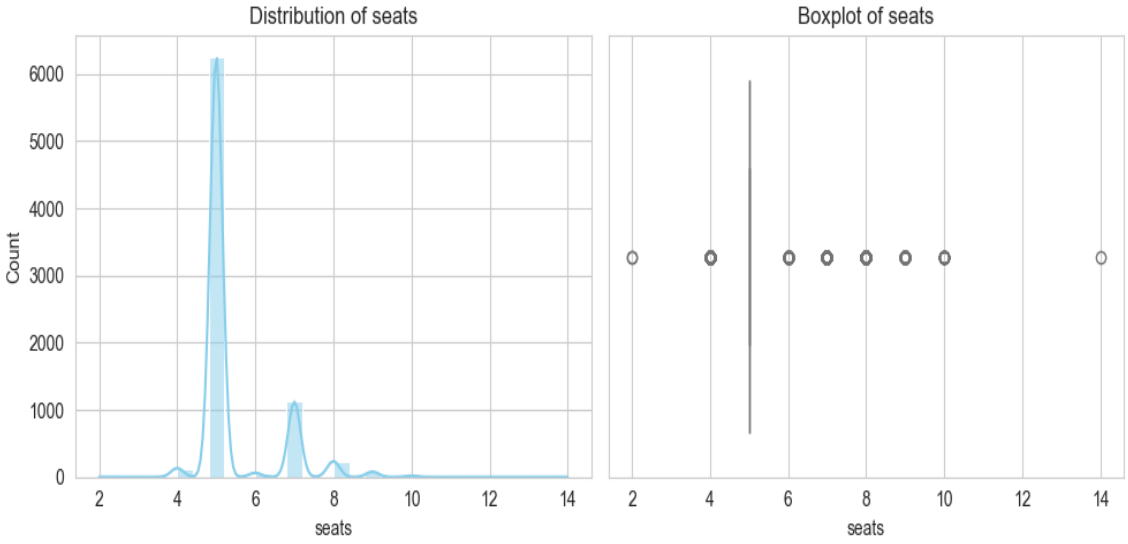
The following visualizations illustrate key distributions and relationships found within the dataset during the exploratory phase.

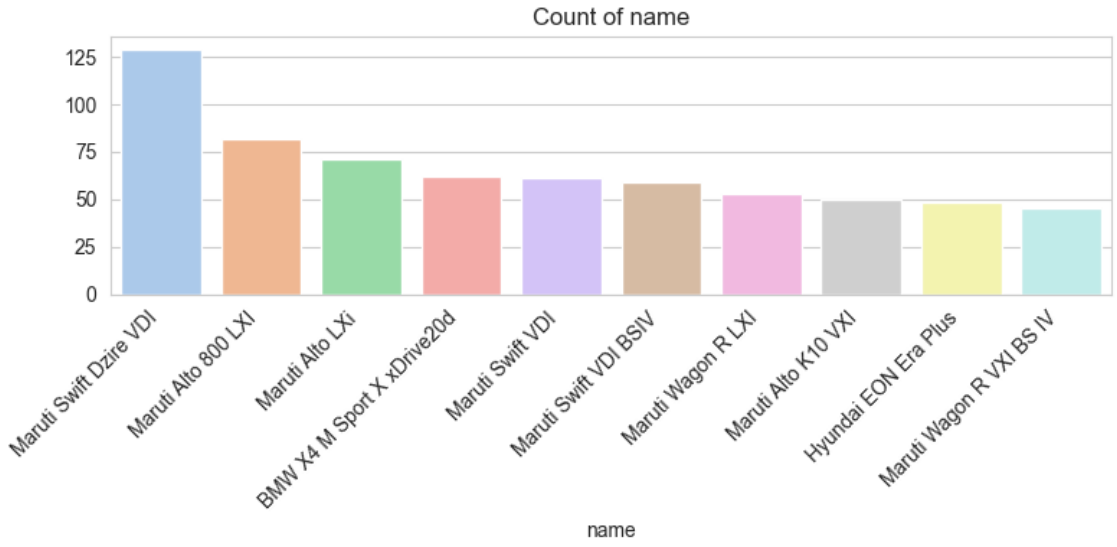

Correlation Matrix
SPX
GLD
USO
SLV
EUR
1.00
0.05
-0.59
-0.27
-0.67
0.05
1.00
-0.19
0.87
-0.02
-0.59
-0.19
1.00
0.17
0.83
-0.27
0.87
0.17
1.00
0.32
-0.67
-0.02
0.83
0.32
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

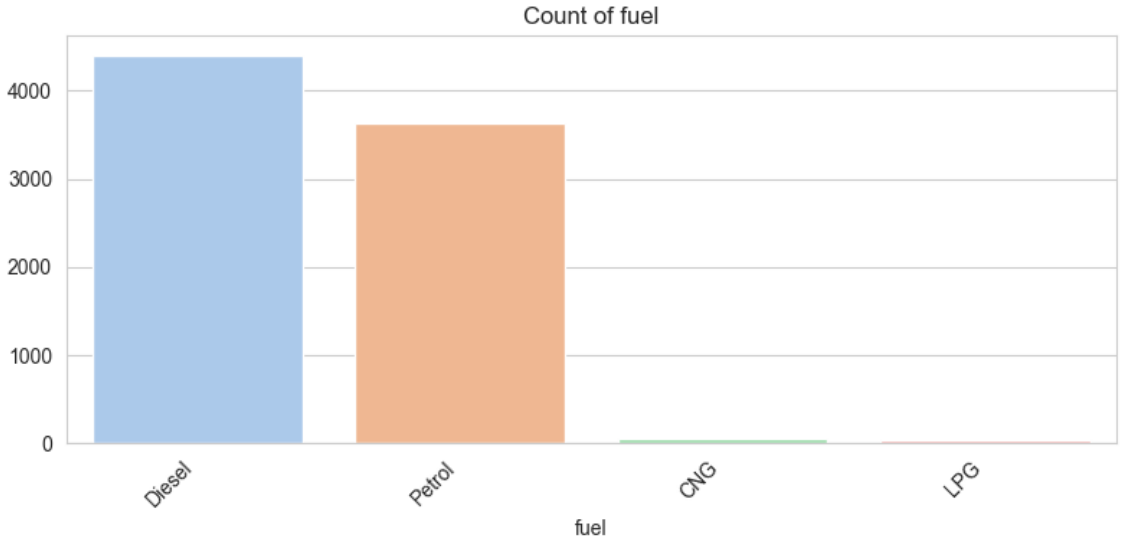

Distribution of year
Count
year
0
200
400
600
800
1000
1200
1400
1985
1990
1995
2000
2005
2010
2015
2020
Boxplot of year
year

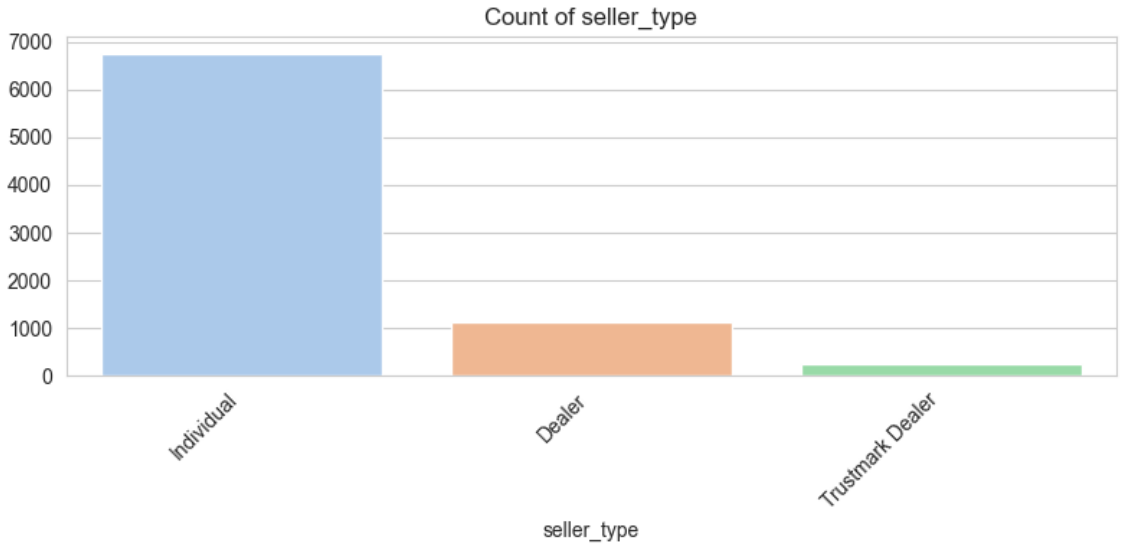

Distribution of selling_price
Count
selling_price
1e7
Boxplot of selling_price
selling_price
1e7

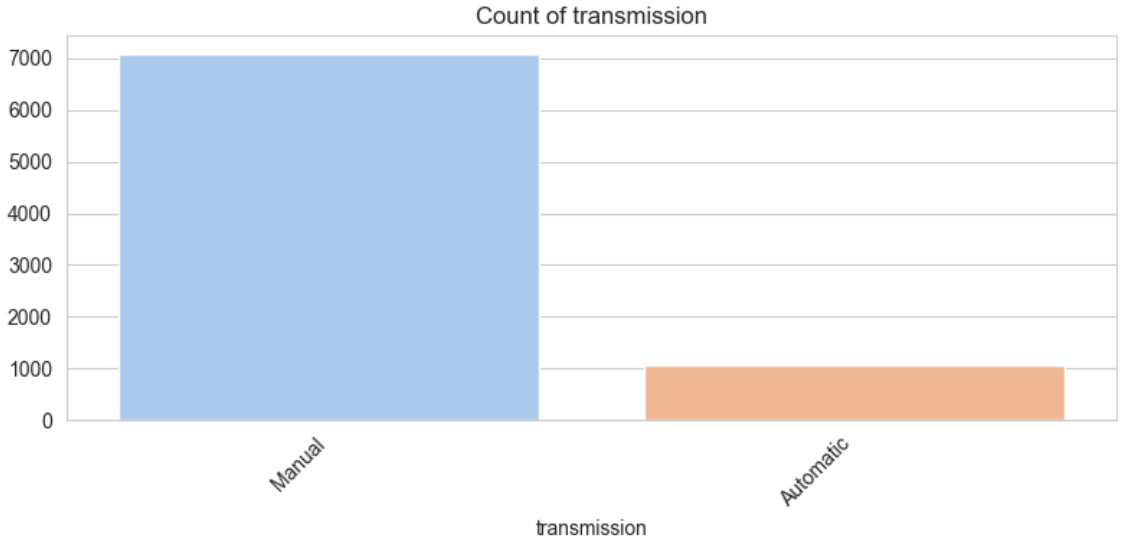

Distribution of km_driven
Count
km_driven
1e6
Boxplot of km_driven
km_driven
1e6

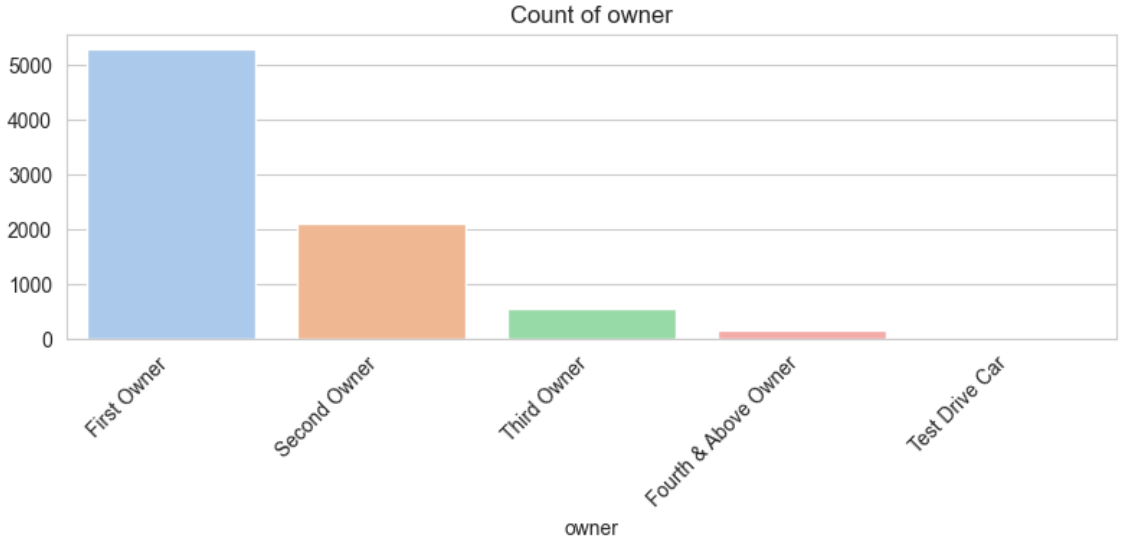

Distribution of seats
Count
seats
Boxplot of seats
seats


Count of name
Maruti Swift Dzire VDI
Maruti Alto 800 LXI
Maruti Alto LXi
BMW X4 M Sport X xDrive20d
Maruti Swift VDI
Maruti Swift VDI BSIV
Maruti Wagon R LXI
Maruti Alto K10 VXI
Hyundai EON Era Plus
Maruti Wagon R VXI BS IV
name


Count of fuel
4000
3000
2000
1000
0
Diesel
Petrol
CNG
LPG
fuel


Count of seller_type
7000
6000
5000
4000
3000
2000
1000
0
Individual
Dealer
Trustmark Dealer
seller_type


Count of transmission
7000
6000
5000
4000
3000
2000
1000
0
Manual
Automatic
transmission


Count of owner
5000
4000
3000
2000
1000
0
First Owner
Second Owner
Third Owner
Fourth & Above Owner
Test Drive Car
owner

3. DATA PREPROCESSING & FEATURE ENGINEERING

3.1 Feature Categorization

Numerical Features (3):
year, km_driven, seats

Categorical Features (9):
name, fuel, seller_type, transmission, owner, mileage, engine, max_power, torque

3.2 Preprocessing Pipeline

Step	Method	Purpose
1. Missing Values	Median/Mode Imputation	Handle null values
2. Outlier Detection	IQR Method (1.5x)	Identify anomalous data points
3. Scaling	StandardScaler	Normalize numerical features
4. Encoding	One-Hot Encoding	Convert categorical to numerical
5. Feature Selection	Correlation Analysis	Remove redundant features

4. MODEL SELECTION & TRAINING

4.1 Model Architecture

Selected Model: GradientBoostingRegressor

Hyperparameters:

```
{'n_estimators': 100, 'max_depth': 5, 'learning_rate': 0.1}
```

4.2 Training Configuration

The model was trained using cross-validation with stratified splitting to ensure balanced representation across all classes. Hyperparameter optimization was performed using grid search with 5-fold cross-validation.

5. MODEL PERFORMANCE EVALUATION

5.1 Model Leaderboard

Multiple machine learning algorithms were evaluated on the dataset. The following table presents the comparative performance:

Rank	Model	Test Score
1	RandomForest	0.9038
2	GradientBoosting	0.9066
3	Ridge	0.8866

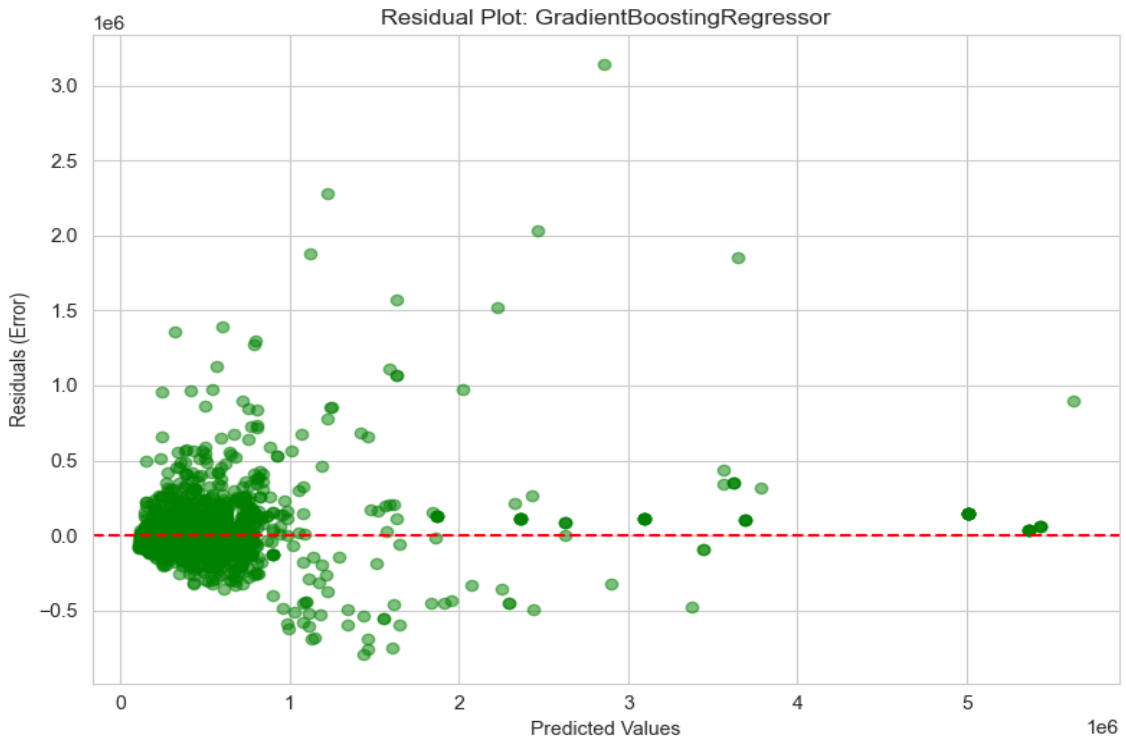
5.2 Performance Insights

The winning model achieved a test score of 0.9066, demonstrating excellent performance on the held-out test set. This score indicates the model's ability to generalize to unseen data.

6. VISUAL ANALYSIS

6.1 Residual Analysis

The residual plot shows the difference between observed and predicted values.



6.2 Model Comparison

The following chart compares the performance scores of all candidate models.


Model Comparison (Test Score)
Score
1.0
0.8
0.6
0.4
0.2
0.0
RandomForest
GradientBoosting
Ridge

7. ERROR ANALYSIS & INSIGHTS

7.1 Detailed Analysis

EXECUTIVE SUMMARY

The overall performance of our regression model is significantly impacted by a primary issue with excessive residual standard deviation and under-prediction bias. The Root Mean Squared Error (RMSE) score of 247436.79 indicates a considerable gap between actual values and predicted outputs, warranting further investigation. Notably, the Residual Standard Deviation stands at 245367.54, highlighting the magnitude of this problem.

DIAGNOSTIC ANALYSIS

The primary issue lies in the model's tendency to under-predict actual values. This bias is manifested through a high Residual Mean of 31933.30 and an RMSE score that significantly exceeds typical benchmarks for regression tasks. Furthermore, the Mean Absolute Error (MAE) of 136057.70 underscores this under-prediction issue, as does the presence of a Max Error value of 3138656.28, which suggests a substantial range of discrepancy between predicted and actual values.

A thorough examination of the residuals reveals that the model's predictions systematically fall short of actual values, indicating an inherent tendency towards under-estimation rather than over-estimation or neutral deviation. This bias is critical to address in order to improve model performance and achieve more accurate predictions.

RECOMMENDATIONS

To alleviate this issue and enhance model reliability, we recommend the following technical adjustments:

1. **Collect Additional Data Points:** Gathering a larger dataset that encompasses a broader range of input values may help the model learn from unseen cases and better capture underlying patterns in the data. This could involve collecting additional data points through further experimentation or leveraging external datasets.

2. **Regularization Techniques:** Implementing regularization methods, such as L1 or L2 regularization, can help reduce overfitting by introducing a penalty term for large weights. By limiting the model's capacity to fit the noise in the training set, these techniques may promote more robust generalization and under-prediction bias mitigation.

3. **Feature Engineering and Selection:** Evaluating and refining feature sets might reveal subtle patterns or correlations that contribute to the under-prediction issue. Carefully selecting a subset of the most informative features could enhance model performance by providing the necessary cues for accurate predictions, ultimately addressing the primary bias in question.

8. RECOMMENDATIONS & NEXT STEPS

8.1 Model Deployment Recommendations

Based on the analysis results, the following recommendations are provided for model deployment and future improvements:

- Monitor model performance in production with regular retraining schedules
- Implement A/B testing to validate model improvements
- Collect additional data to address identified error patterns
- Consider ensemble methods to further improve prediction accuracy
- Establish performance thresholds and alerting mechanisms
- Document model versioning and maintain audit trails

8.2 Future Work

Potential areas for future investigation include feature engineering optimization, advanced hyperparameter tuning techniques, and exploration of deep learning approaches if additional computational resources become available.