Forecasting Auto Loan Originations: A Data-Driven Approach for the Bank

Huyen Le

August 2023

Outlines

1. Introduction
2. Data Processing
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Selection and Training
6. Model Comparison
7. Recommendations

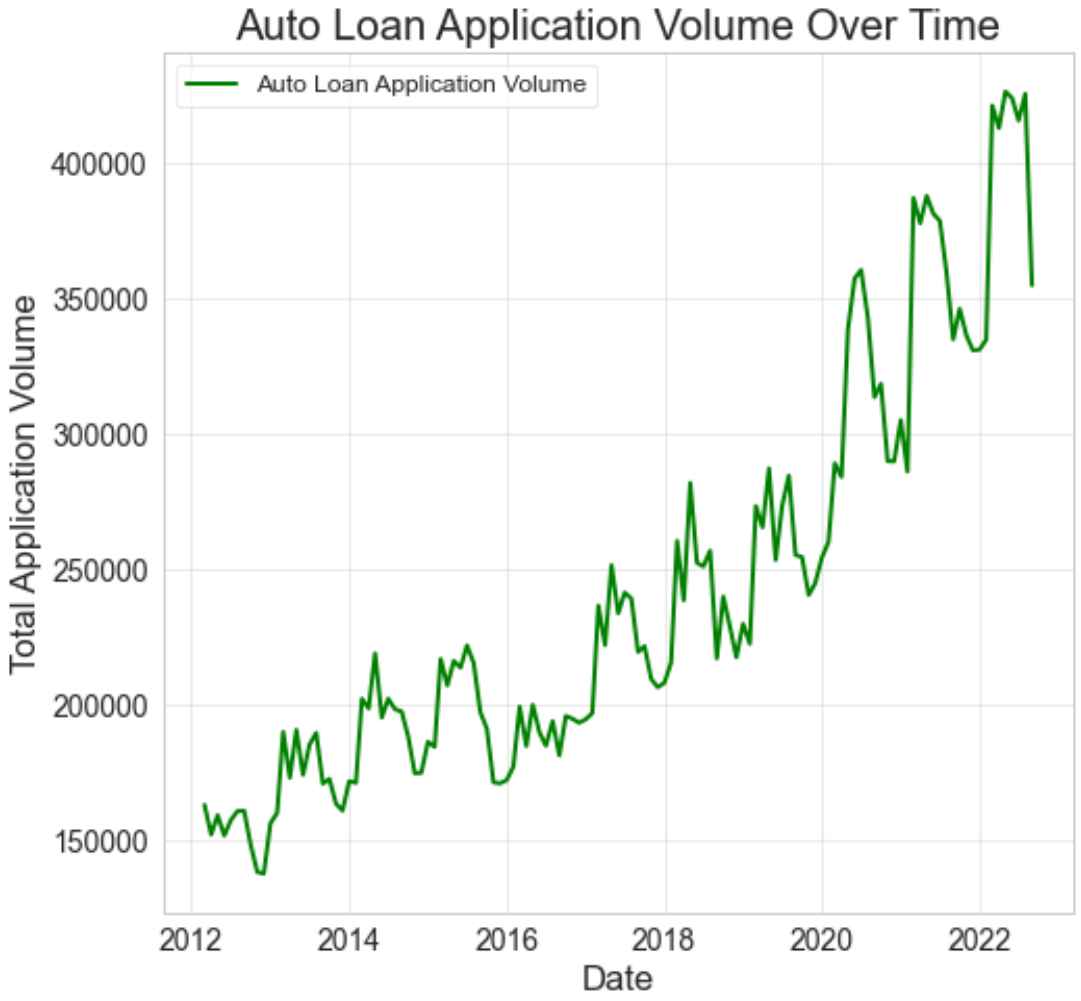
Introduction

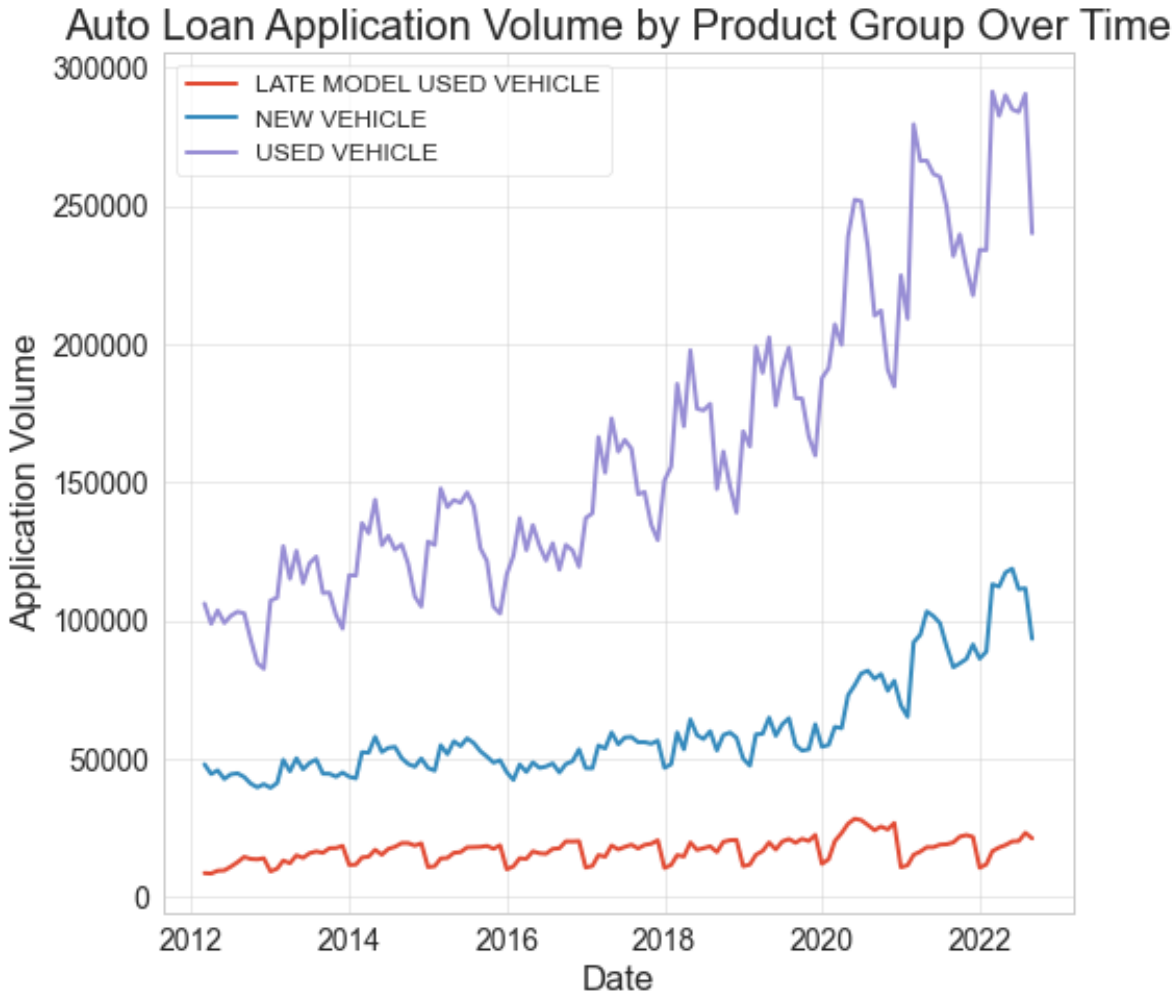
- *Bank's Offer*: Provides auto loans to its diverse client
- *Business Need*: Anticipating loan demands is crucial due to fluctuating market dynamics and evolving customer behaviors.
- *Role of Analytics*: The analytics team leverages data to forecast trends, ensuring preparedness.
- Challenges:
  - Diverse Factors: Impacted by both internal metrics and external elements.
  - Data Integration: Essential to merge varied data sources for insights.
  - Dynamic Market: Economic shifts and changing preferences complicate forecasting.

Data Processing

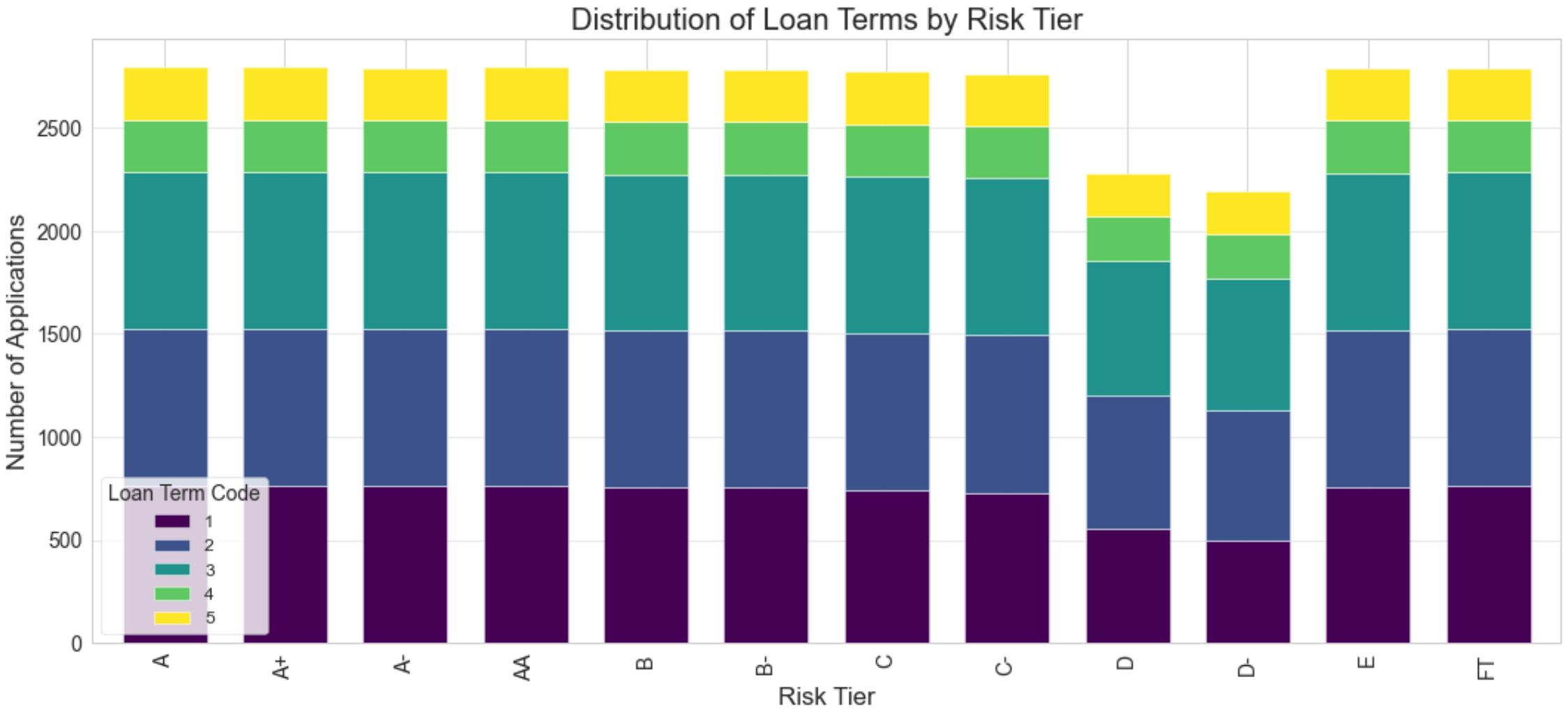
- The dataset to predict auto loan originations contains:
  - Total membership
  - Application volume (Group of product, pre-approve, risk tier, term code, discount and total application volume)
  - Competitor rates
  - Macroeconomic indicators
- Data processing: Merged data resulted in:
  - 12 missing values in "Pre-approve“, and 220 in "Total membership"
  - Missing values were imputed using the mean.
  - Outliers detected using the Interquartile Range and replaced with the median.
  - Drop 1,127 duplicated observations
  - Label encoding and one-hot encoding applied for categorical variables.
- Dataset Size: 32,344 observations spanning 10 years.

Data Overview

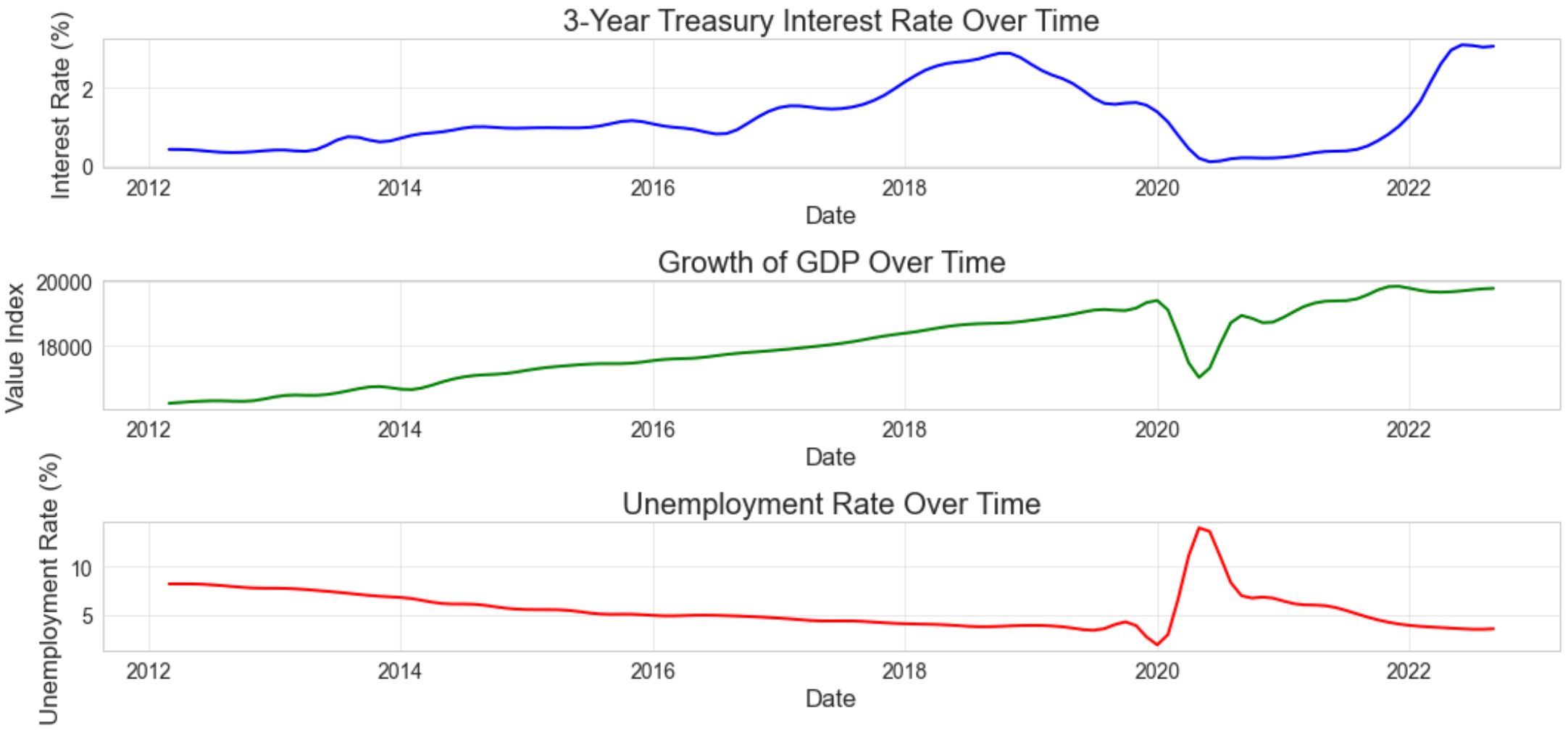




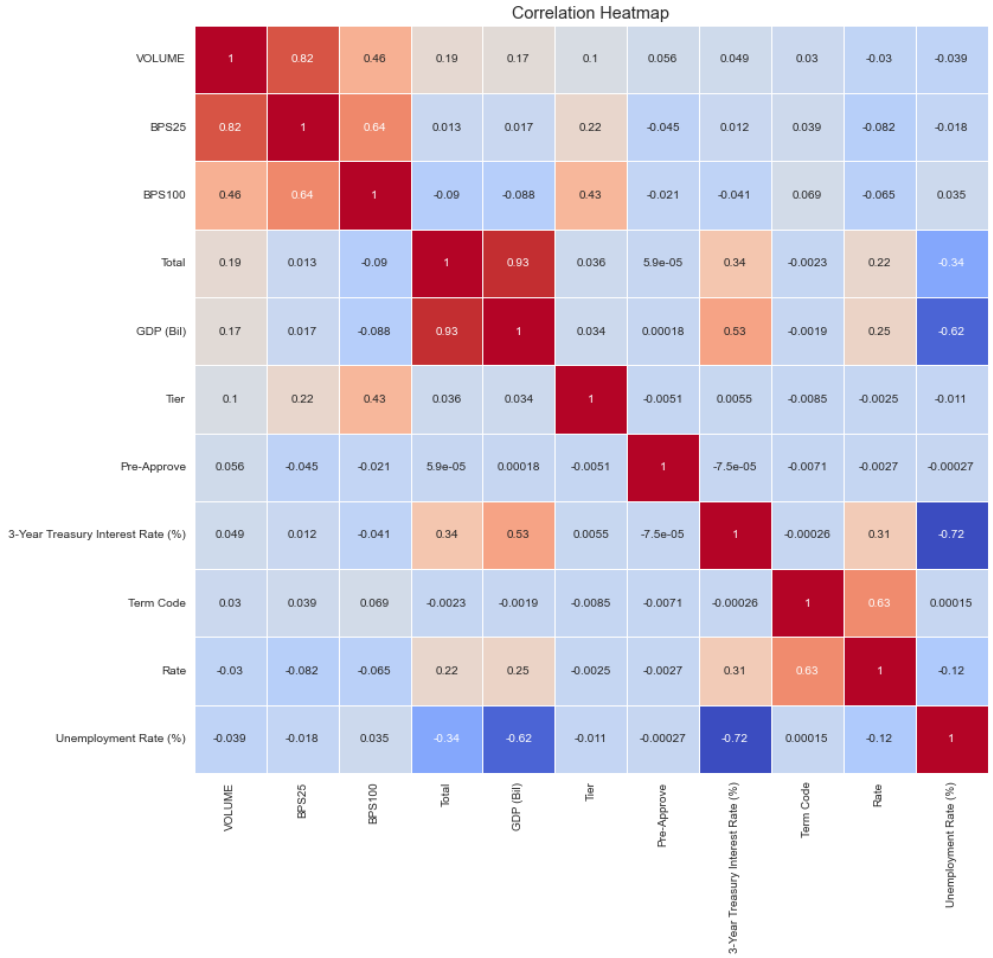
Data Overview



Data Overview



Feature engineering



	coef	std err	t	P>|t|
const	-409.7928	88.872	-4.611	0.000
PREAPPROVE	94.9068	3.203	29.630	0.000
TIER	-5.5179	0.559	-9.880	0.000
BPS25	15.9242	0.126	126.084	0.000
BPS100	24.5023	1.263	19.399	0.000
TERM_CODE	46.5765	2.015	23.118	0.000
Total	0.0023	0.000	18.894	0.000
Rate	-8575.5901	340.849	-25.160	0.000
3-Year Treasury Interest Rate (%)	47.8462	3.475	13.767	0.000
GDP (Bil)	0.0116	0.005	2.136	0.033
Unemployment Rate (%)	34.4323	2.244	15.345	0.000
PRODUCT_GROUP_NEW VEHICLE	36.9167	4.205	8.779	0.000
PRODUCT_GROUP_USED VEHICLE	311.9972	5.556	56.155	0.000

Adjusted R-squared: 0.68

Model Selection and Training

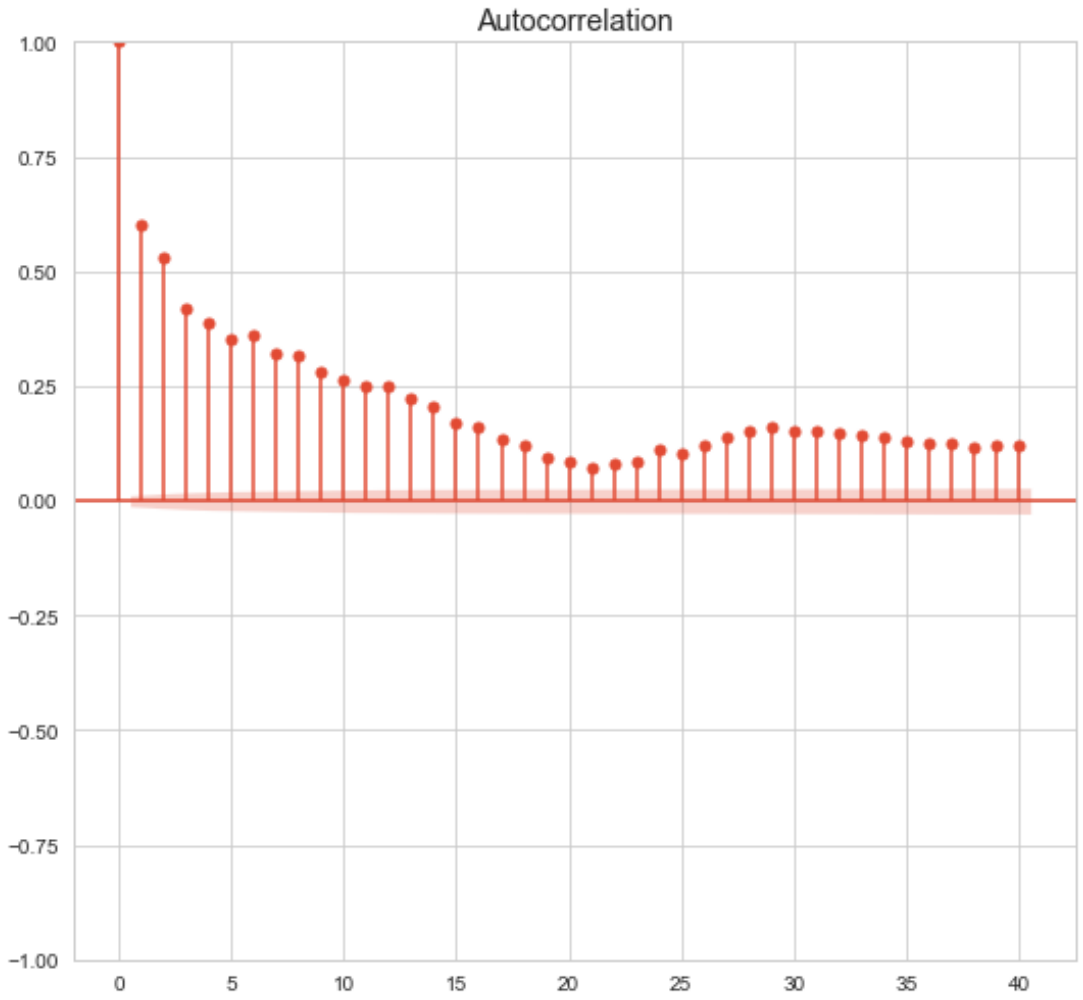
- Train-test split: set aside 20% as test data, 80% training data
- Apply for the most popular models
  - Linear regression model
  - Time-series model (ARIMA)
  - Decision tree
  - Random Forest
- Decision made:
  - Based on the result of Mean absolute error (MAE), Mean squared error (MSE) and Adjusted-R squared.

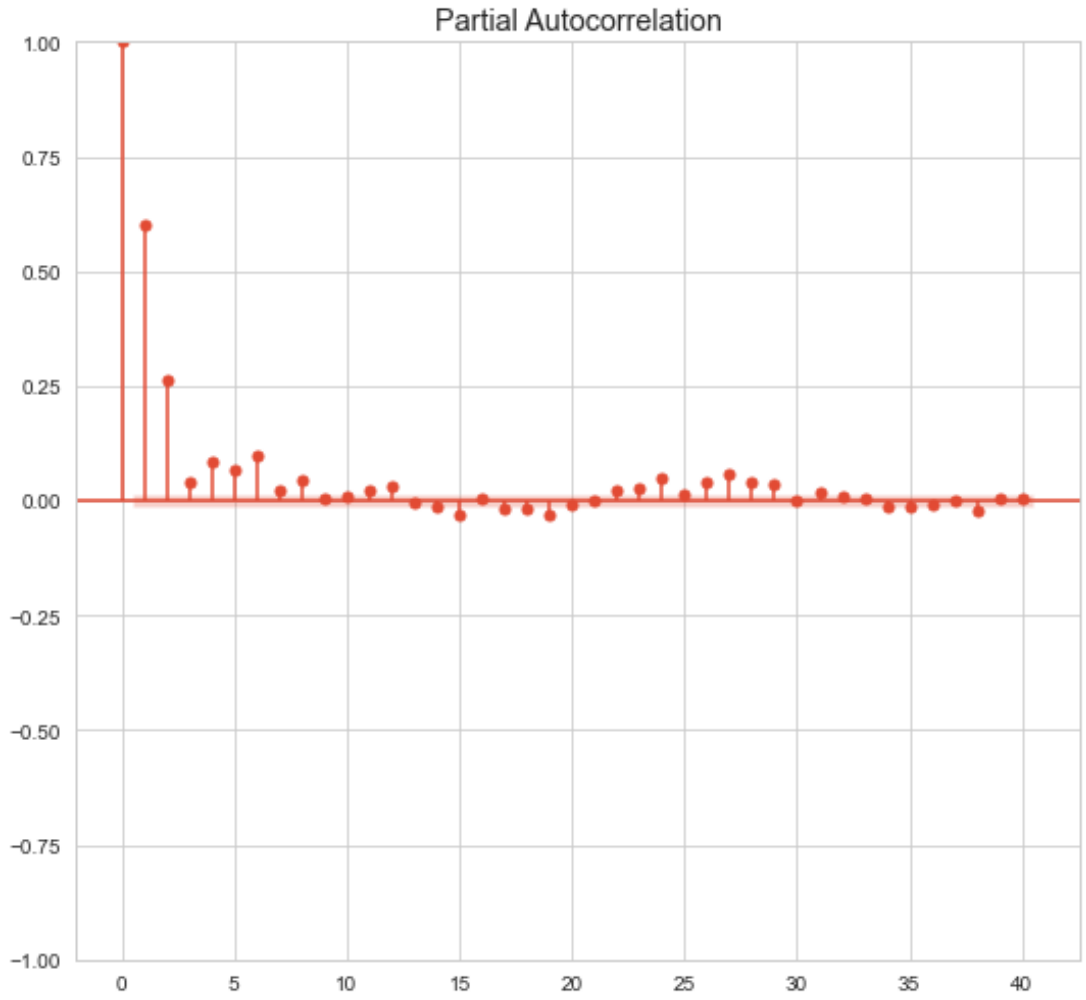
Linear Regression Model

- Mean Absolute Error (MAE): 2.39×10^{-12}
- Mean Squared Error (MSE): 8.4×10^{-23}
- R^2 : 1.0
- Conclusion: Perfect fit => overfitting?
  - Inspect the coefficients to find the significant impact of lags of Volume => Drop lags
  - Solution: Drop lags => MAE: 202.9, MSE:83352.6, R^2:0.68

ARIMA Model

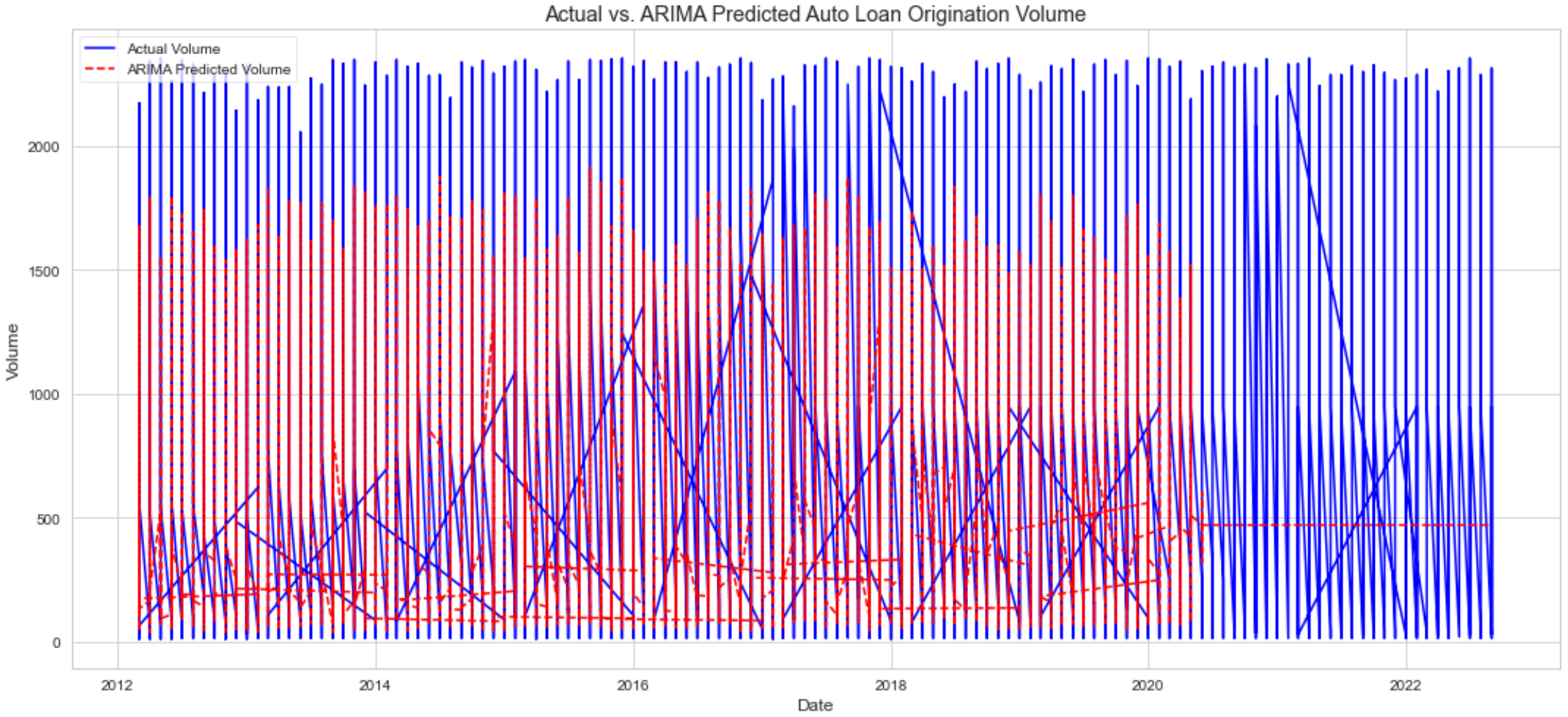
(p,d,q)=(2,1,2)





ARIMA Model

MAE:436
MSE: 299167
R^2 :-0.11



Machine Learning Model

- **Decision Tree Regressor:**
  - Mean Absolute Error (MAE): 131.814
  - Mean Squared Error (MSE): 55742.38
  - R^2 : 0.79

- **Random Forest Regressor:**
  - Mean Absolute Error (MAE): 107.90
  - Mean Squared Error (MSE): 38876.17
  - R^2 : 0.85

Model Comparison

Model	MAE	MSE	R^2
Linear regression model	202.9	83352.6	0.68
ARIMA	436	299167	-0.11
Decision Tree	131.814	55742.38	0.79
Random Forest	107.90	38876.17	0.85

Model Robustness

- Stress-testing:
  - Purpose: To evaluate the model's stability and reliability under extreme or unusual conditions.
  - Outcome for Linear Regression: A 5% decrease in R-squared when introduced with outliers.
- Sensitivity analysis:
  - Purpose: To assess the influence of changes in independent variables on a specific dependent variable.
  - Result for Linear Regression: By omitting lag features, the R-squared value decreased by 32%.

Recommendation

- Collect more historical data on auto loan originations
- Evaluate the events and factors that influence auto loans
- Try different models and techniques to make the result more accurate
  - Use k-fold cross section techniques
  - Linear regression model: use Ridge regression or Lasso regression to avoid overfitting
  - Time-series model: consider different methods, or drop the insignificant lag features, add the significant interaction features
  - Random Forest/Decision Tree: Find the best hyperparameter tuning