Cross-Sell Prediction

Ali Malenchik

Background

- Cross-selling involves marketing new products to existing customers. Cross selling increases revenue.

- Machine learning can be used to predict which customers will be responsive to a cross-sell attempt, allowing companies to use marketing resources efficiently.

Problem Statement

Can a machine learning algorithm accurately predict whether a health insurance customer would be receptive to cross-sell attempt for vehicle insurance?

Scope & Assumptions

- This project will use encoding in order to prepare categorical features for modeling. This may result in a highly dimensional dataset.
- This project uses data pipelines to perform hyperparameter tuning.

Methods

Data Source

The dataset used for this project is a collection of information on cross-sell outcomes for health insurance policyholders.

Each record contains information about a unique customer and his or her response to the attempt to cross-sell vehicle insurance.

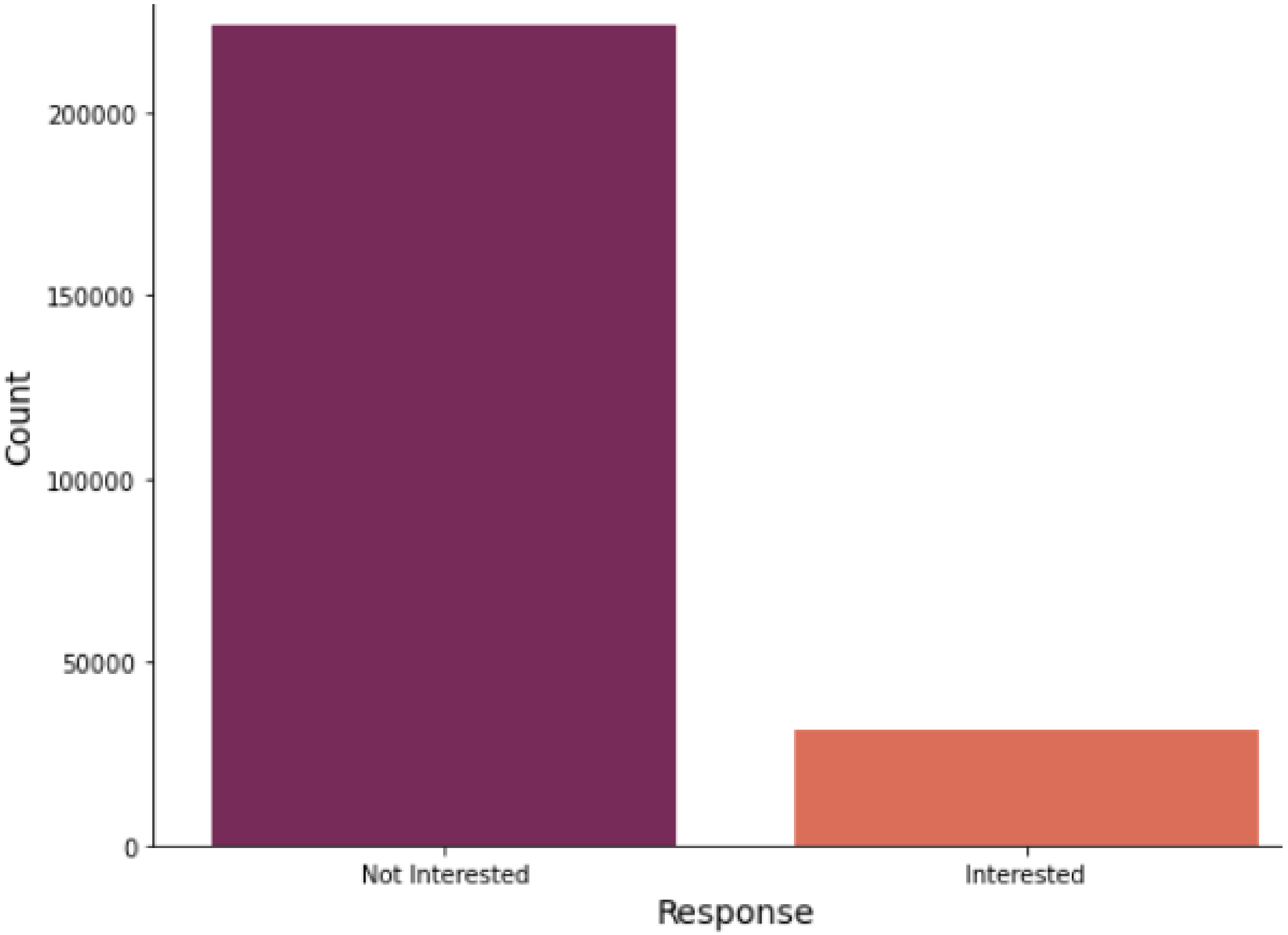
- Sourced from Kaggle
- CSV format
- 12 columns and 381,109 rows
- Binary target variable is "Response" (1 for interested, 0 for not interested)

Data Import & Cleansing

- Region_Code and Policy_Sales_Channel updated to string and decimals removed.
- ID attribute removed
- No null values
- Split into 33% test, 67% train

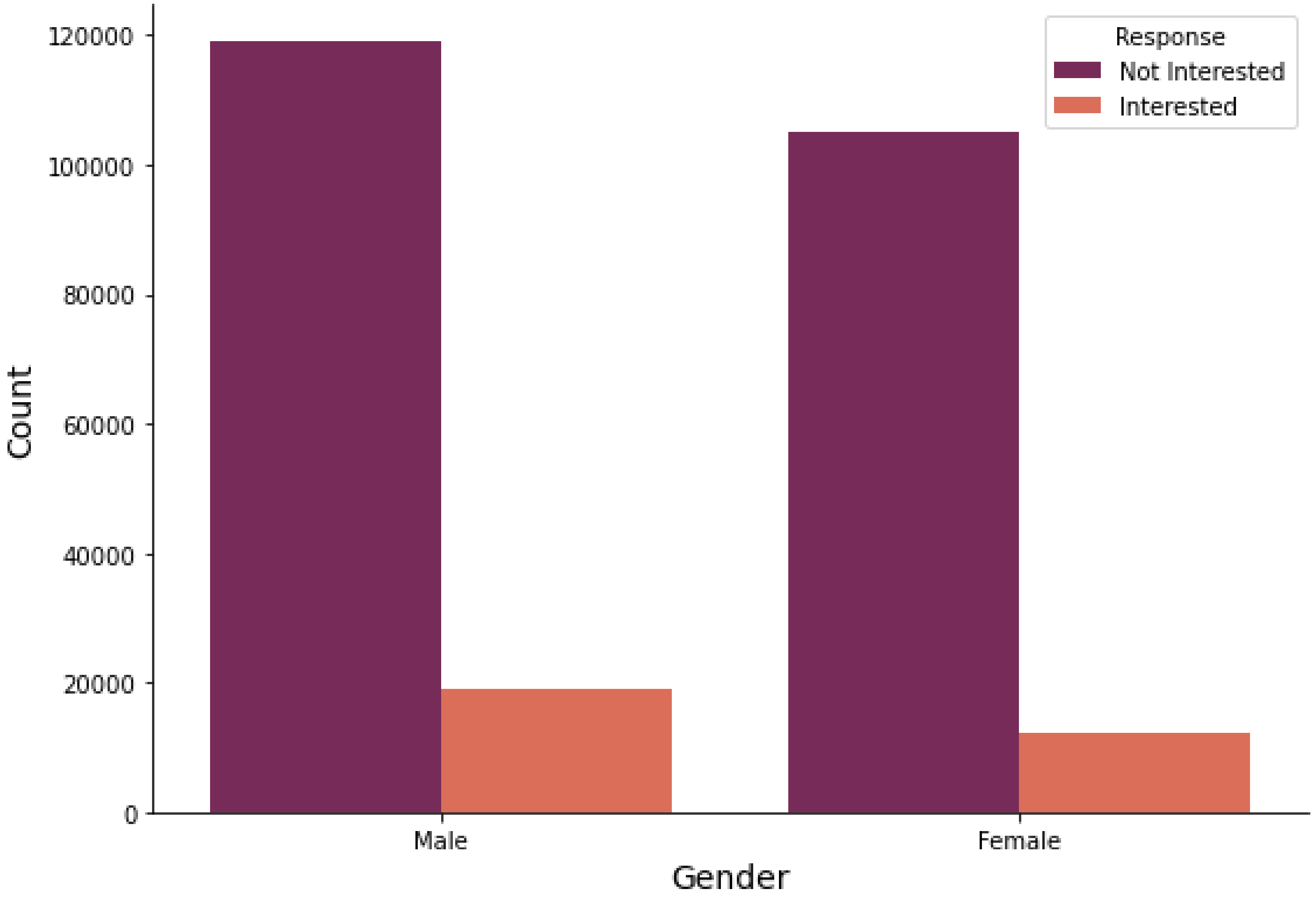
Exploratory Analysis

Count Plot of Policyholder Response to Cross-Sell Attempt



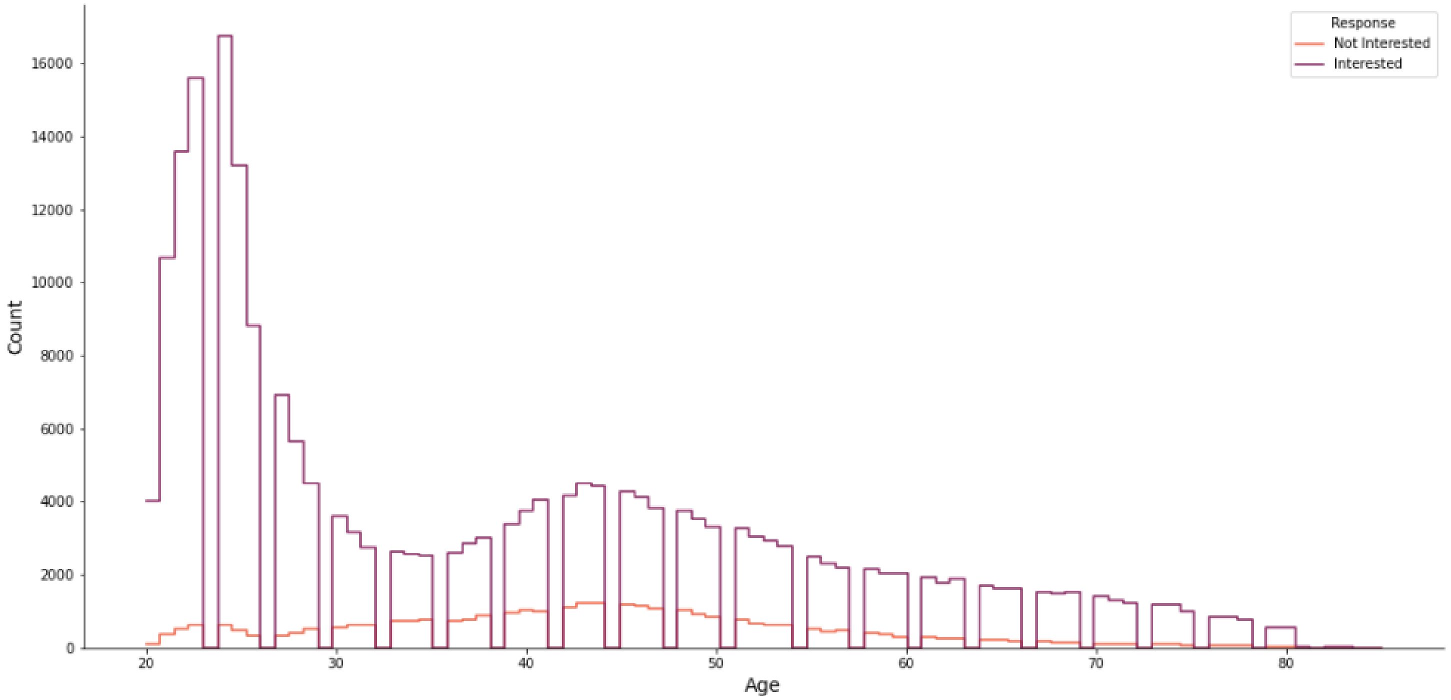
Exploratory Analysis

Grouped Count Plot of Policyholder Response by Gender



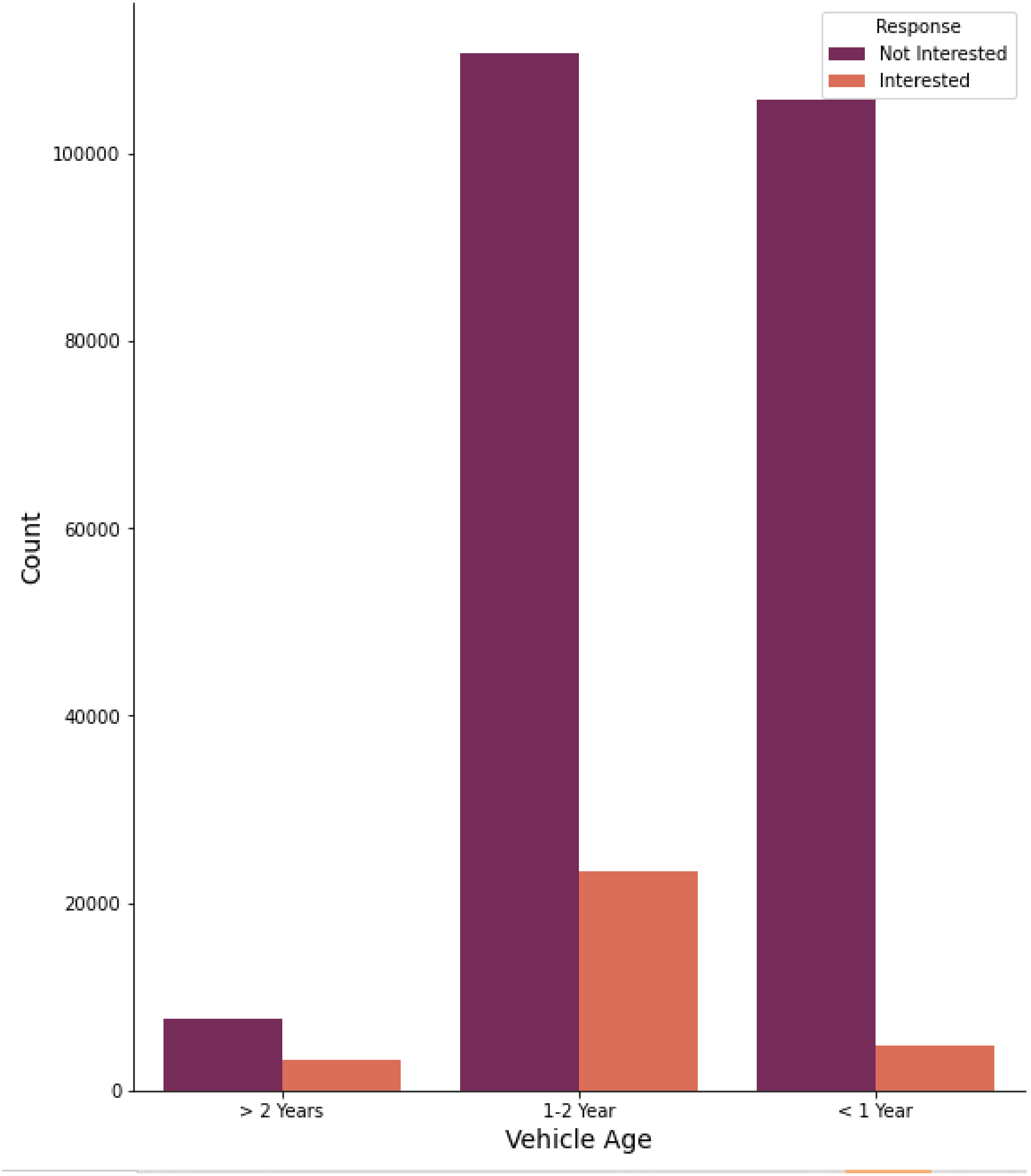
Exploratory Analysis

Grouped Step Plot of Policyholder Response by Age



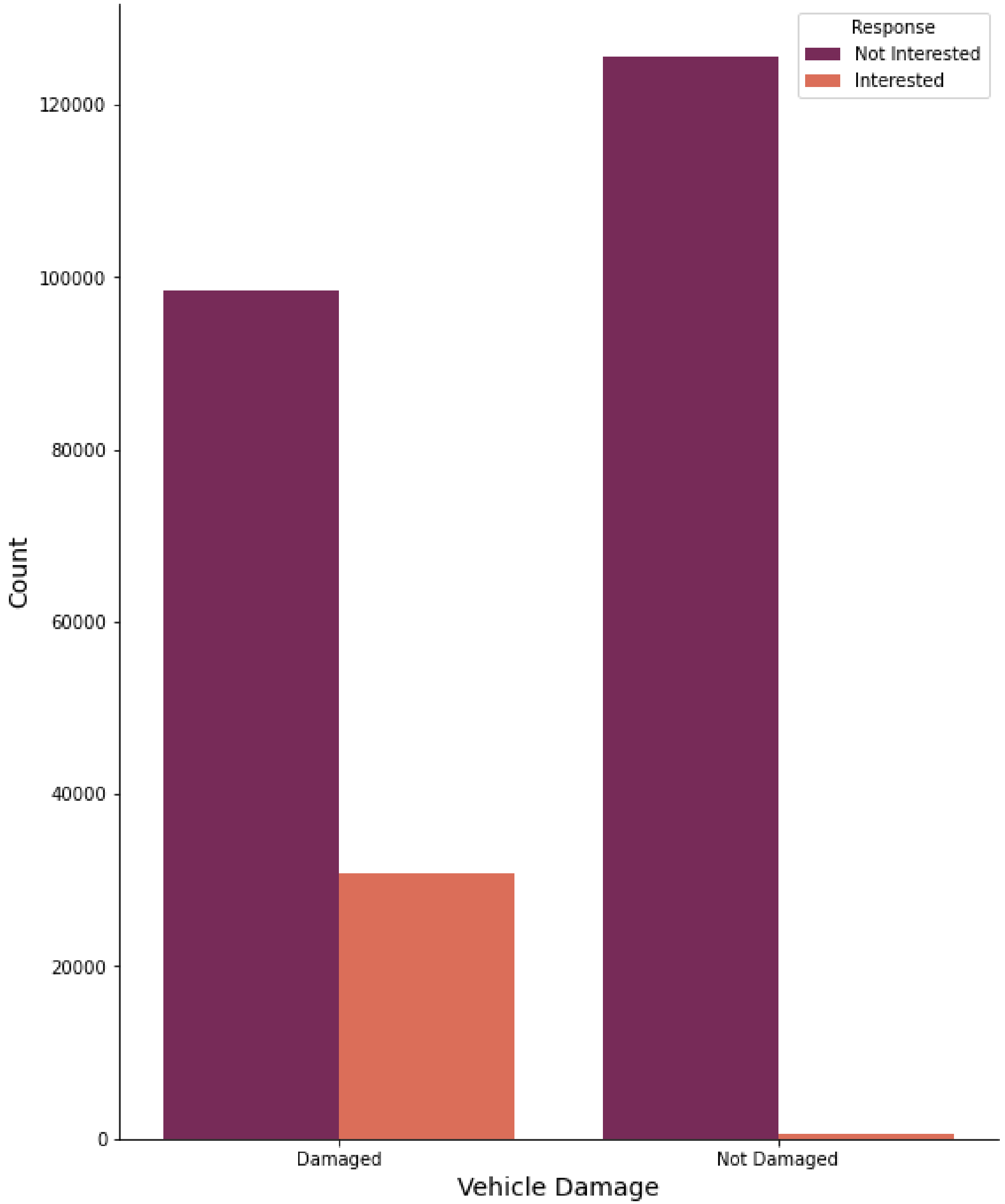
Exploratory Analysis

Grouped Count Plot of Policyholder Response by Vehicle Age



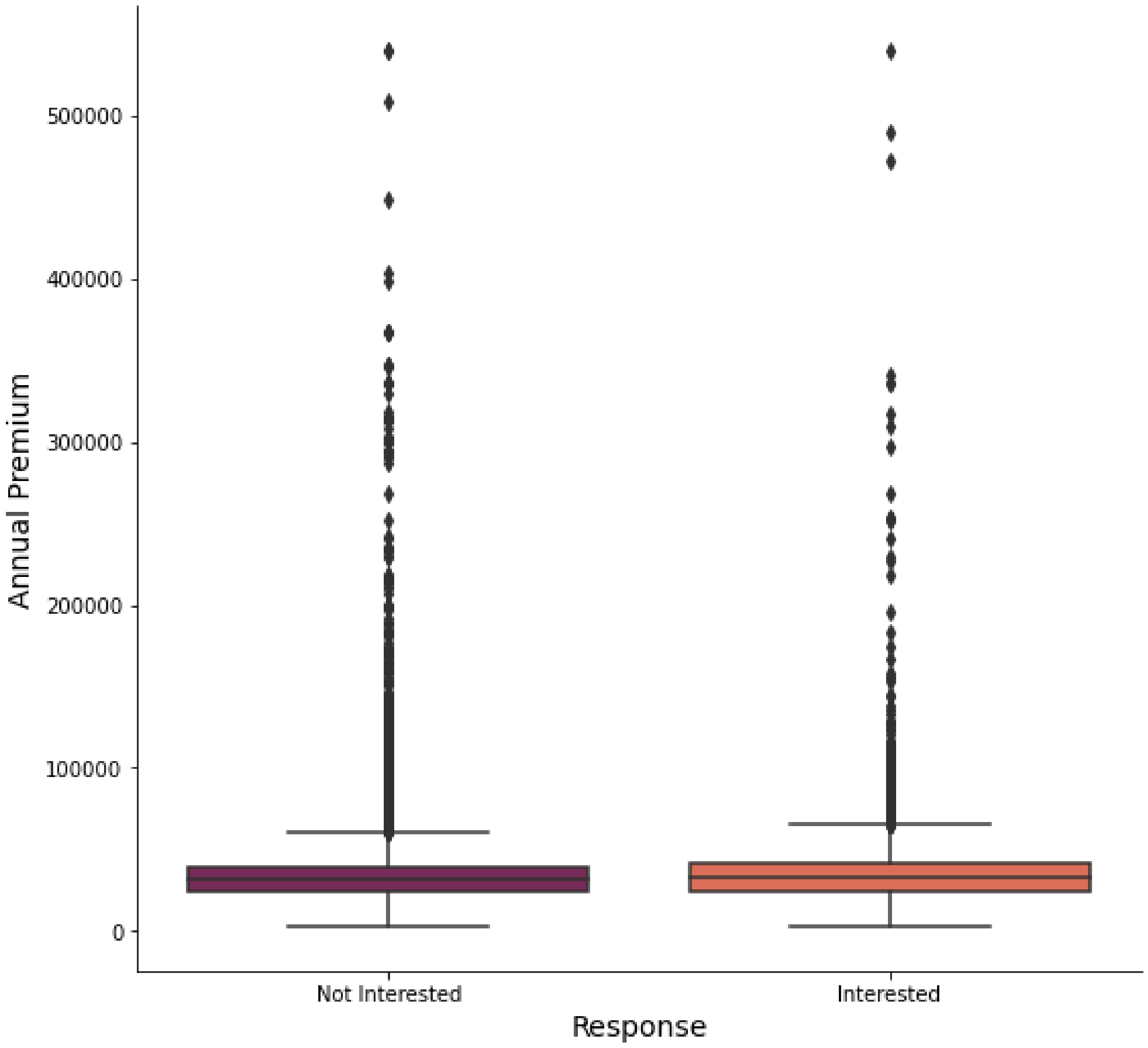
Exploratory Analysis

Grouped Count Plot of Policyholder Response by Vehicle Damage Status



Exploratory Analysis

Grouped Box Plot of Policyholder Response by Annual Premium



Feature Selection

1. Categorical values encoded, resulting in 214 features
   - Vehicle_Damage
   - Gender (Only one column included)
   - Region_Code
   - Vehicle_Age
   - Policy_Sales_Channel

2. Relationship with target variable calculated and features with weak relationship with target variable dropped, resulting in 33 features

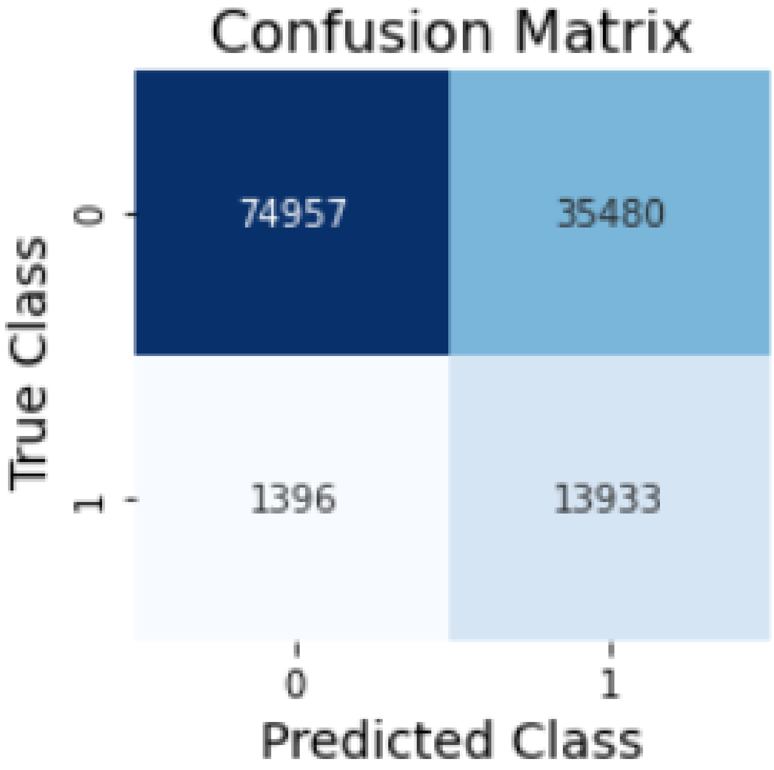
Model Deployment

Pipelines created for 3 models & hyperparameter tuning performed using grid search with 5-fold cross-validation

- Logistic Regression (ROC AUC .837)
- Random Forest (ROC AUC .846)

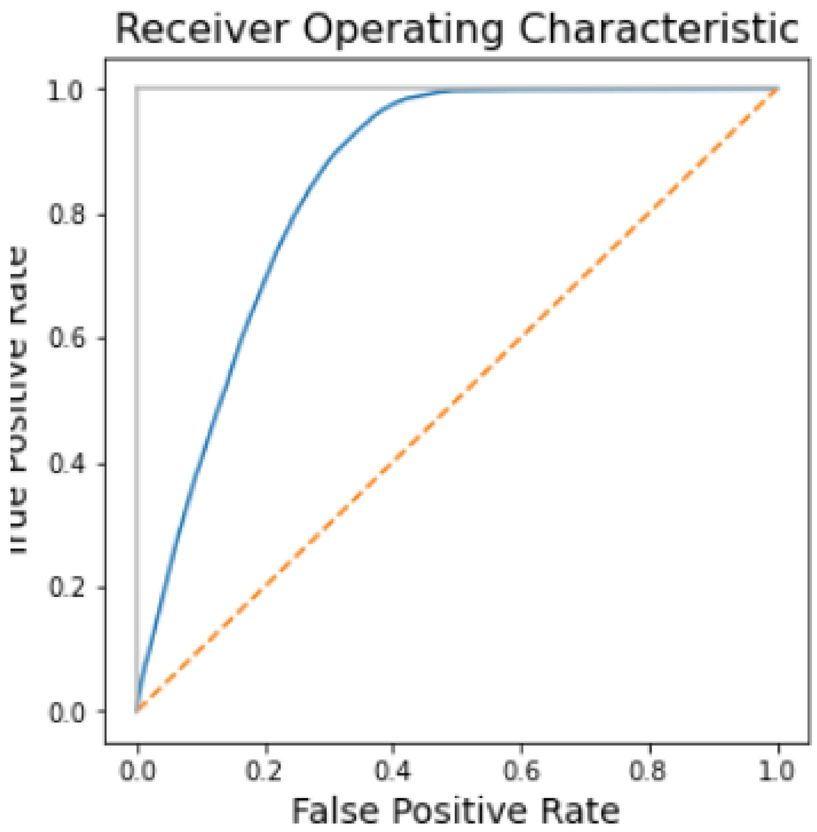
Results

The best performing random forest model was evaluated.



Confusion Matrix

True Class \ Predicted Class	0	1
0	74957	35480
1	1396	13933



Receiver Operating Characteristic

Conclusion

The efficiency and performance of the random forest model can likely be improved with better feature selection and enhanced parameter tuning.