Insurance Cross-Sell Prediction

Ali Malenchik

Bellevue University

DSC680: Applied Data Science

Prof. Catherine Williams

October 24, 2021

Abstract

Cross-selling is a valuable marketing technique that involves encouraging existing customers to purchase or enroll in another product or service within the same company. Using machine learning to predict which customers will be responsive to another product allows businesses to save time and money by allocating marketing resources where they are most effective.

The goal of this project was to develop a classification algorithm that accurately predicts whether a health insurance customer would be receptive to cross-marketing of vehicle insurance. Information collected about 380,000 insurance policyholders was used to identify patterns and characteristics of a customer who responds “yes” to a cross-sell attempt. This data was encoded and used to perform feature selection, model training, and hyperparameter tuning to determine the best performing model between Random Forest and Logistic Regression.

Based on the results, Random Forest was found to be better performing, although the performance could likely be improved with better feature selection techniques and further parameter tuning. Due to the high dimensionality of the data, model training was inefficient and limited.

Introduction

Background

Cross-selling is the concept of marketing additional products to existing customers in order to earn additional revenue. Cross-selling is a common practice in financial industries, insurance industries, and more. It provides an easy way to grow the business and gain profits. It also creates loyalty among customers – the more products they use from a brand, the more loyal they are to that brand and the longer the customer retention. However, attempting to cross-sell to every customer would be inefficient and ineffective. Not all consumers would benefit from or find value in signing up for another product from a company. Marketing to these customers is not only a waste of time and resources, but it could also result in a negative experience for the customer. Using machine learning to predict which customers are responsive to cross-selling allows businesses to effectively target communications and optimize marketing strategies.

Problem Statement

Can machine learning accurately predict whether a health insurance customer would be responsive to a 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 utilizes 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 381,109 health insurance policyholders found on Kaggle. The dataset is provided in CSV format and contains twelve columns:

- ID: A unique ID to identify each customer
- Gender: Male or Female
- Age: Age of the customer in years
- Driving_License:
    - 0: Customer does not have a driver's license
    - 1: Customer does have a driver's license
- Region_Code: Unique code for the customer's region of residence
- Previously_Insured
    - 1: Customer already has vehicle insurance
    - 0: Customer does not have vehicle insurance
- Vehicle_Age: Age of the vehicle in years
- Vehicle_Damage
    - 1: Customer's vehicle has been damaged in the past
    - 0: Customer's vehicle has not been damaged in the past
- Annual_Premium: The amount the customer pays for health insurance each year
- Policy_Sales_Channel
    - Anonymized code for the method of outreach (i.e. different agents, mail, phone, in person, etc.)

- Vintage
    - Number of days the customer has been associated with the company
- Response
    - 1: Customer is interested in vehicle insurance
    - 0: Customer is not interested in vehicle insurance

The target variable for prediction is “Response”, which indicates whether the health insurance policyholder responded as interested in vehicle insurance from the same company.

Data Import & Cleansing

Data was imported into a data frame using Pandas’ read_csv function. At the time of import, the data frame contained 12 columns and 381,109 rows. Each row contains information about a unique customer and his or her response to the cross-sell attempt. At this time some light cleansing was performed. Two attributes, Region_Code and Policy_Sales_Channel, were interpreted as float data type due to the nature of the values. However, since they are to be interpreted as categorical data, they were updated to string and decimal suffixes (“.0”) were removed. The id attribute was removed, as it does not provide any relevant information that could be used in prediction. At this time the data was also analyzed for null values, and none were observed. Finally, the data was split into 33% test, 67% train using sklearn’s train_test_split function. After splitting the data frame, the training dataset contained 255,343 rows and the test dataset contained 125,766 rows.

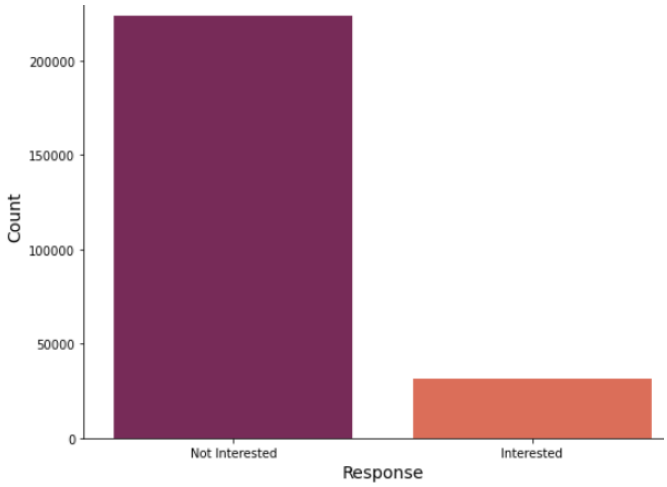
Exploratory Analysis

During exploratory data analysis, multiple plots were created. First, a count plot of policyholder response to the cross-sell attempt was created using Seaborn’s countplot. The

training dataset demonstrated an imbalanced target class, with over 200k customers being not interested in vehicle insurance.

Figure 1

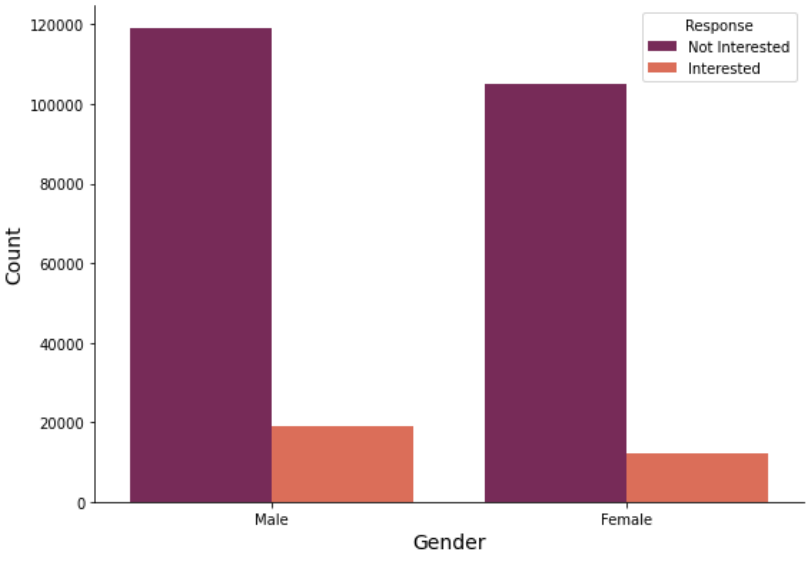
Count Plot of Policyholder Response to Cross-Sell Attempt



Next, the distribution of policyholder response was compared for male and female policyholders. Interestingly, a higher proportion of male policyholders seemed to respond positively to the cross-sell attempt.

Figure 2

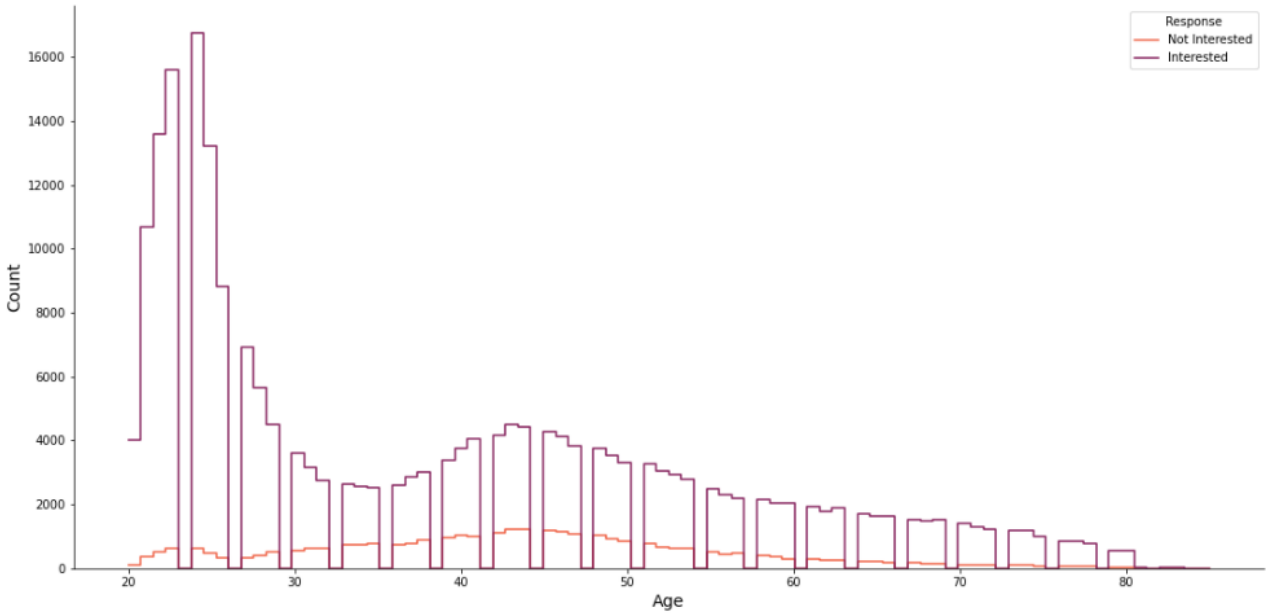
Grouped Count Plot of Policyholder Response by Gender



Multiple visualizations were created to analyze the policyholder's age. First, a histogram of age was created using Seaborn's histplot. From the histogram we can gather that the most frequent age for policyholders is in the mid-twenties. There is another spike around early to mid-fourties, and then a steady decline until the mid-eighties age. In order to compare the policyholder response for each age, a grouped step plot was created.

Figure 3

Grouped Step Plot of Policyholder Response by Age



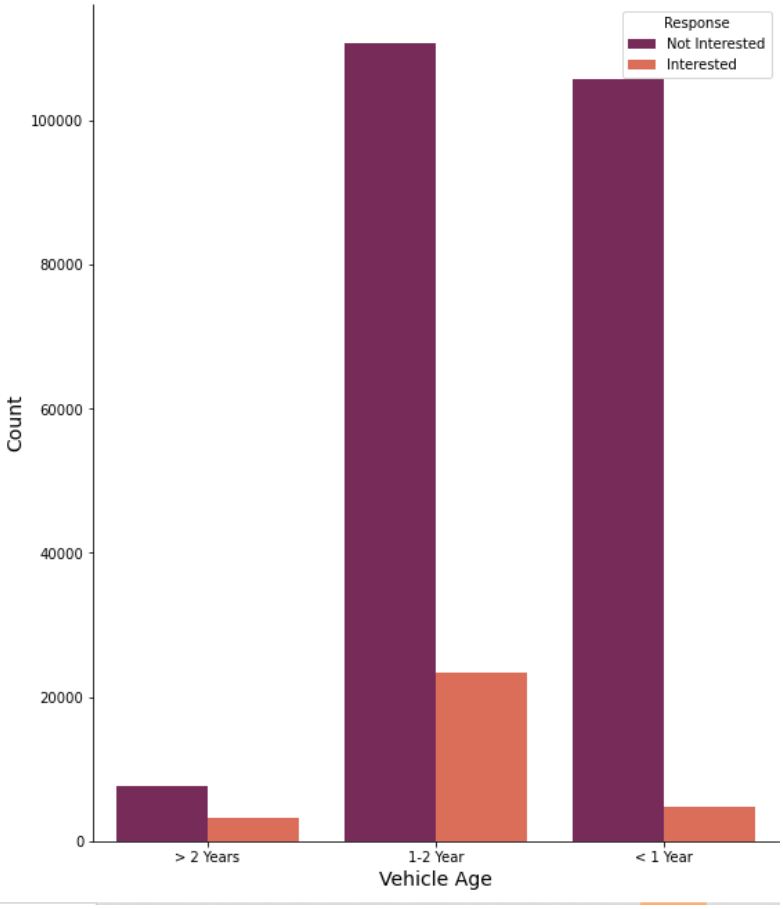
The step plot demonstrates that despite the spike in the mid-twenties age, the proportion of younger policyholders that are responsive to cross-selling is very low compared to other ages. Additionally, from the box plot comparing the same information, we can see that the IQR for those interested in the cross-sell product is much higher and narrower in age range – around 35 to 52 – compared to those not interested in the cross-sell product – around 24 to 49. The mean age of non-responsive policyholders is around 34, compared to non-responsive policyholders at around 43.

Unsurprisingly, grouped box plots of the policyholder response by drivers license status showed that customers without a driver's license showed little to no interest in the vehicle insurance. Similarly, those who already have vehicle insurance at another company showed very little responsiveness to cross-selling. This would indicate that marketing is better aimed toward policyholders with a driver's license but no existing vehicle insurance.

The grouped count plot of response by vehicle age showed that a higher proportion of policyholders with older cars (>2 years old) showed interest in the cross-sell. Those with newer cars (<1 year old) showed very little interest in comparison.

Figure 4

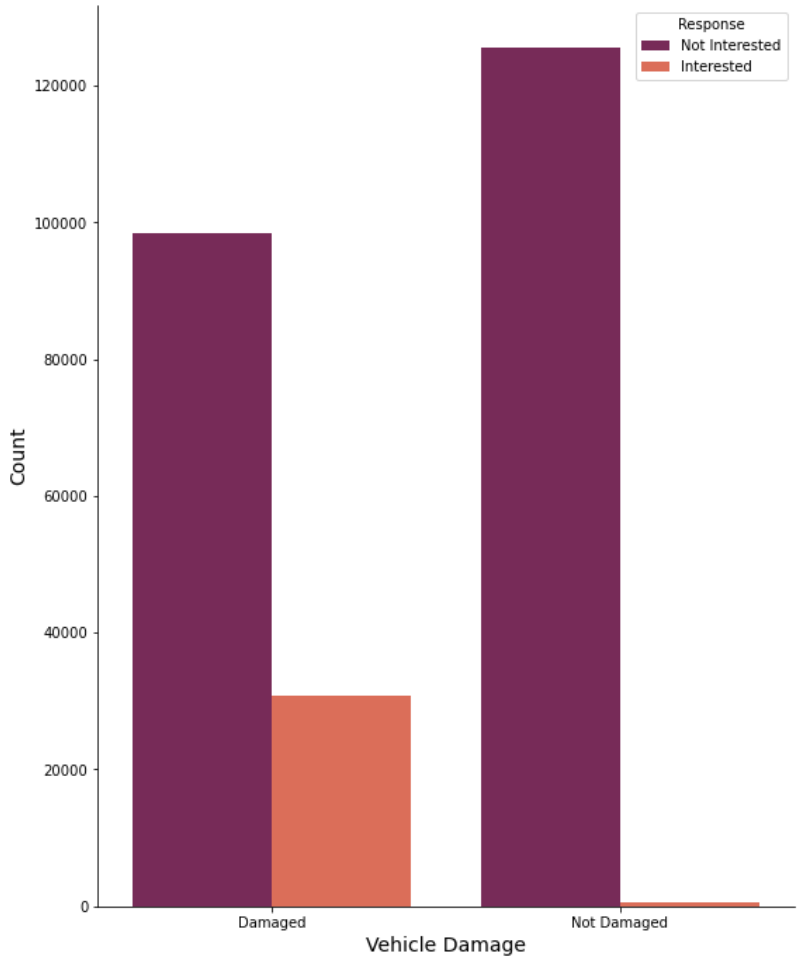
Grouped Count Plot of Policyholder Response by Vehicle Age



In the grouped box plot of policyholder response by vehicle damage status, we can see that approximately 1/3 of policyholders with vehicle damage were interested in the cross-sell, whereas almost none of the policyholders without vehicle damage were interested.

Figure 5

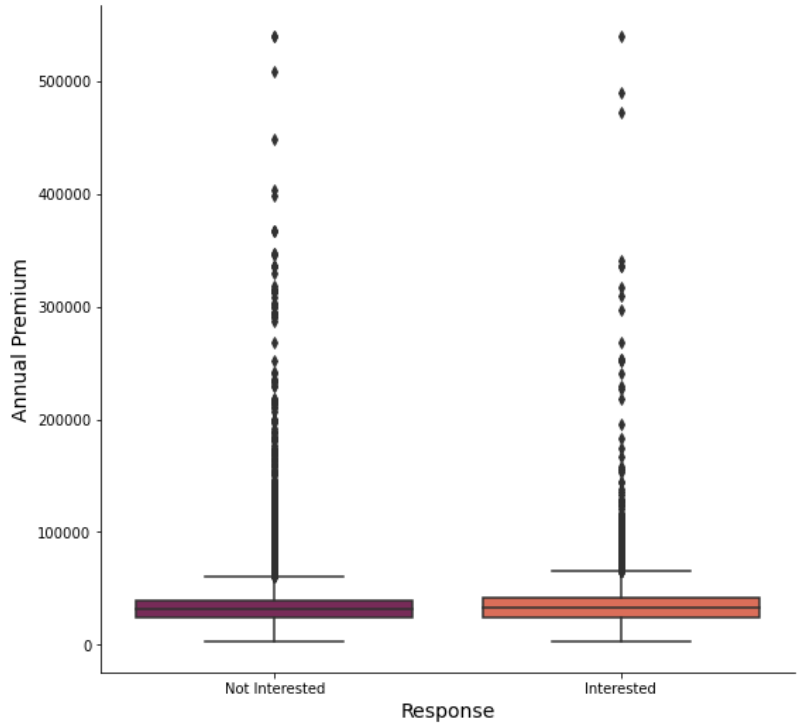
Grouped Count Plot of Policyholder Response by Vehicle Damage Status



Multiple visualizations were created to analyze the policyholder annual premium. The histogram of annual premium showed that the majority of premiums fell below 100,000, with the median around 30,000. The grouped step plot showed that the distribution for both non-responsive and responsive policyholders was similar. The grouped box plot indicated almost identical ranges and IQRS for both groups.

Figure 6

Grouped Box Plot of Policyholder Response by Annual Premium



The histogram and box plots of policyholder days with company showed a relatively uniform spread, with almost identical distributions for responsive and non-responsive policyholders.

Multivariate analysis showed interesting relationships between the policyholder age and their vehicle age. Younger policyholders in their 20's tended to have newer cars (<1 year old), whereas policyholders in their 30's and up tended to have vehicles 1-2 years old. The number of policyholders having cars over 2 years old was relatively low across all ages.

Based on multivariate scatterplots (annual premium vs policyholder age, policyholder age vs. days with the company, days with the company vs. annual premium), very little could be learned as the distributions were uniform and no observable patterns could be seen.

Feature Selection

In order to prepare the data for feature selection, categorical values were encoded. The Vehicle_Damage attribute was converted from Yes/No values to 1/0 values. The remaining categorical variables (Gender, Region_Code, Vehicle_Age, Policy_Sales_Channel) were encoded using category_encoders' OneHotEncoder function. Since having two colums related to gender is redundant, Gender_Female was dropped from the data frame. After this, the data frame consisted of 214 columns. A correlation heatmap was created to identify very highly correlated features. None were observed.

Next, independent variables were split from the dependent variable (Response). Relationships between the features and target were determined using sklearn's mutual_info_classif function. The features having a value of less than .001 were dropped. After this step, only 33 columns remained.

Model Deployment

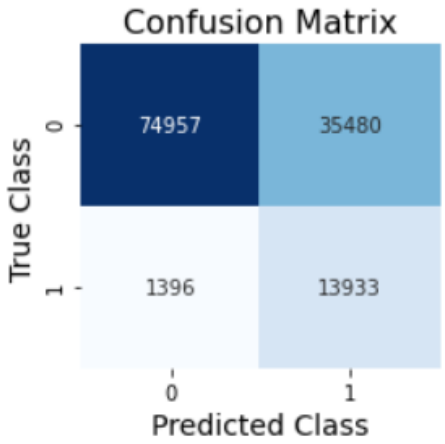
During model training, pipelines were created for logistic regression and random forest models. SelectKBest was chosen as the selector for all models. Hyperparameters were chosen based on the parameters available for each. A grid search with 5-fold cross-validation was performed using roc_auc scoring metric. The best performance for the logistic regression model resulted in a ROC AUC score of .837, whereas the best performance for random forest resulted in a score of .846. Thus, random forest will be considered the best performing model.

Results

The best performing model was evaluated using a confusion matrix. The confusion matrix showed a large number of false positives (35,480).

Figure 7

Confusion Matrix of Random Forest Model



A classification report was also computed, which demonstrated a low precision score for the positive class (.28), but high precision for the negative class (.98) and high recall for the positive class (.91). Finally, the Receiver Operating Characteristic (ROC) curve was plotted to visualize the results of evaluation.

Figure 8

Plot of the Receiver Operating Characteristic Curve for Random Forest Model

Discussion

Conclusion

Cross-selling is a valuable tool that allows businesses to market products to an already existing client base. Using machine learning to predict which customers are responsive to cross-selling is more efficient and cost-effective, reducing marketing waste and increasing revenue. Because historical information can be gathered on the responsiveness of customers along with other customer characteristics, cross-selling is a prime candidate for training a classification model to predict success of a cross-selling attempt.

Limitations/Challenges

Unfortunately, because of the high dimensionality of the dataset after encoding, the efficiency of the model training was very poor which resulted in limited parameters selected for tuning. The grid search step in SVM model training was particularly time consuming and it was decided not to proceed with this method. As a result, the performance of the “best” model is not ideal.

Next Steps

The next steps for this project would be to explore potentially reducing the size of the dataset by eliminating more features or performing under-sampling, and determining the resulting impact on model performance. This would help make the model training more efficient. I would also explore how SVM performance specifically could be optimized by narrowing down the parameters in model training.