Final Report: Insurance Cross-Sell Prediction

Submitted to:

Dr. Casey Bennett

School of Computing and Digital Media

DePaul University

Report Prepared By:

Natalia Guzman

November 18, 2024

Abstract

The goal of this project is to use machine learning models to analyze the likelihood of customers from a healthcare insurance company to buy their vehicle insurance. To achieve this, Random Forest and Gradient Boosting were applied on the dataset after performing SMOTE for balancing the target variable and feature selection. The results demonstrated an outstanding performance of Random Forest with an accuracy of 90% and AUC of 0.97 which highlights the advantages of this model to address cross-selling prediction tasks.

1. Introduction

In the insurance industry, companies are constantly exploring innovative methods to enhance profitability and customer retention. One promising approach is cross-selling—promoting additional products to existing customers. This strategy not only strengthens customer relationships but also maximizes the value of each customer, reducing the need to acquire new ones. However, the challenge in cross-selling lies in effectively identifying customers who are likely to respond positively to such offers. In the context of vehicle insurance, predictive modeling provides an opportunity to achieve targeted cross-selling by leveraging historical and demographic data. By applying machine learning, insurers can predict which customers may show interest in additional products based on their existing profiles.

This project focuses on developing a predictive model to determine customer interest in cross-selling insurance policies for a specific dataset from a health insurance company. The dataset includes demographic and behavioral features such as age, insurance history, and vehicle damage records, all of which are relevant to the insurance sector. The primary objective is to utilize machine learning models, specifically Random Forest and Gradient Boosting, to create a reliable predictive framework for customer interest. In addition to maximizing prediction accuracy, the project emphasizes the need for model interpretability, enabling insights into which factors most significantly influence customer interest in cross-selling offers. The findings aim to provide insurers with a tool to make targeted, data-informed marketing decisions, optimizing resource allocation while enhancing customer engagement and satisfaction.

2. Literature Review

The automobile and vehicle insurance remains one of the most aggressive industries, increasingly using predictive analytics as well as customer segmentation to provide optimal services and sales. One of the most effective growth strategies of cross-selling is a firm that markets the additional products to its existing customers, which increases the customer's value proposition as well as engagement and potential customer lifetime value [1].

Cross-selling in insurance gained interest because it would help increase profitability without needing to acquire new customers. Historically, insurers have used policyholder data to identify who is likely to be sold another product. This allows for targeted marketing and product recommendations. Cross-selling also helps enhance retention because the establishment of a multiple-product relationship supports loyalty better. For instance, concerning vehicle insurance, the history of a customer and demographic can be utilized as a determinant for health or life products by an insurer, hence increasing the value of a customer's lifetime [2].

The literature presents that cross-selling is not very easy as the recommendations need to follow the requirements of the customer rather than being an annoyance, or this may cause him to disengage, as he becomes irritated. Insurers can therefore mitigate such challenges while still optimizing their cross-selling strategy through predictive analytics [3].

Machine Learning Applications in the Insurance Domain

Machine learning has revolutionized insurance, especially in underwriting, claims prediction, fraud detection, and customer retention strategies [4]. ML algorithms take on vast amounts of data that project patterns and insights found to be instrumental in determining some decisions.

ML models can segment customers on grounds of factors influential to their purchasing behaviors, and help in cross-selling. From the given research, it can be seen that the performance of an ML algorithm is better than the statistical technique regarding attrition prediction of a customer, and estimation of the lifetime value of policyholders in an insurance company, which serves as a base for cross-selling opportunity identification [1].

The Random Forest algorithm is one ensemble learning technique that forms multiple decision trees and merges them together to overcome overfitting and maximize accuracy. Random Forest has been shown to perform well for applications in insurance if the data is complex and noisy, like customer behavior patterns [5]. For instance, in Duan et al. (2022), Random Forest is made to illustrate its proper and efficient identification of high-risk customers for cross-selling assistance through safer and more reliable leads for additional insurance products.

In cross-selling environments, Random Forest enables an insurance company to find a set of predictive features such as customers' demographics, types of policies, and historical claims. This means developing a customized approach toward the recommendation of suitable products for a particular customer segment, which will thereby enhance the effectiveness of cross-selling campaigns [2].

Another ensemble approach is Gradient Boosting, which works successively to improve predictions based on the residual errors of previous models, so even in imbalanced datasets - common in insurance - 'attaining maximum possible accuracy' [6]. Mavundla (2024) showed how effective Gradient Boosting is in customer segmentation for an insurance company as it detected subtle patterns that would help to differentiate between high-value customers in successful cross-selling.

A significant advantage of Gradient Boosting over Random Forest is its ability to pick up strong features from the data that would be useful in informing insurers about tailoring customized offerings for their customers [6]. This algorithm's accuracy and interpretability make it suitable for tasks such as targeting customers in cross-selling campaigns where small precision gains translate into large monetary returns [2].

Many research pieces have focused on the usage of predictive modeling in cross-selling in the insurance sector. To showcase this, for instance, Broby (2019) demonstrated the usefulness of predictive models based on customer data to estimate the probability of policyholders buying new products.

Saxena, S. (2024) undertook a different study where they compared Random Forest with Gradient Boosting in a vehicle insurance product cross-sell predictive model. They found that, while the two algorithms are nearly equivalent at competing, Gradient Boosting offered about 1% more precision than Random Forest in what would prove paramount in averting attrition of customers by undiplomatic recommendation.

The literature shows that techniques of machine learning like Random Forest and Gradient Boosting are methods that effectively enhance cross-selling activities in the insurance sector. Such algorithms improve not only predictive accuracy but also support segmentation in a customer selection aligned with individual needs, maximizing the satisfaction and lifetime value for customers. Given the increased exploration of insurers toward data-driven solutions, the application of predictive modeling in the cross-selling activity will be more extensive to enhance profitability and customer loyalty.

3. Methodology

3.1 Data

The dataset used for this project was obtained from the website Kaggle. It contains information about 381,109 insurance policyholders from a health insurance company and has a total of 12 features. Among the numerical features you can find information about a customer's age, their region code, the amount of money they have to pay for premium per year, the code for the channel used to reach the customer, and the number of days the customer has been associated with the insurance company. On the other hand, the categorical variables are the customer's ID, gender, ownership of driver's license, if they already have vehicle insurance, the age of their vehicle (it is classified into 3 categories which are less than a year, between 1-2, or more than 2 years), and history of damage to the vehicle.

3.2 Data Preprocessing and Exploratory Analysis

While exploring the data, I checked for missing or duplicated values, and there were none. Then, I decided to analyze the distribution of the numerical features. As seen in Figure 1, the feature 'Age' has a right-skewed distribution with a high concentration around young ages, 'Region_Code' has multiple peaks suggesting that some areas are more represented in the dataset than others, 'Annual_Premium' is heavily right-skewed with most values concentrated on the lower end, 'Policy_Sales_Channel' has a multimodal distribution with

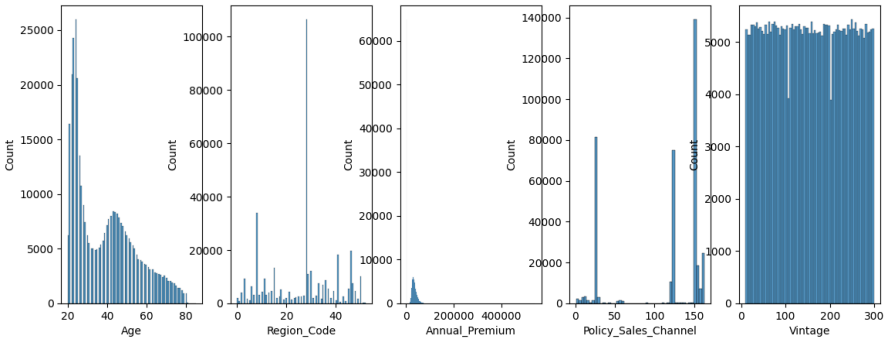

Figure 1. Histograms of numerical features

2-3 channels being more common than the others, and ‘Vintage’ shows a uniform distribution. On the other hand, the categorical variables show a fairly balanced distribution

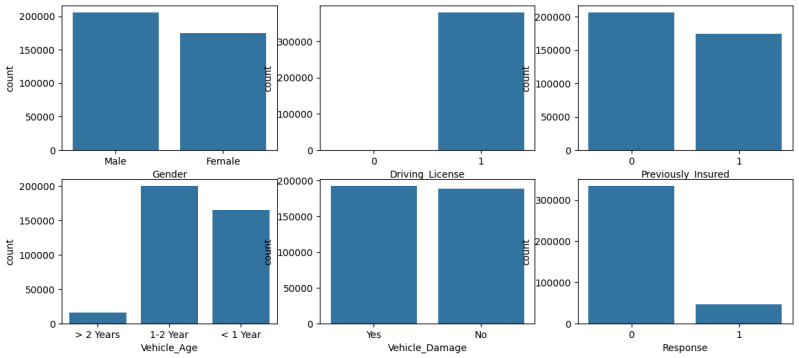

Figure 2. Bar Plots for categorical features

between males and females, people with or without previous insurance, and with or without vehicle damage history. However, we can also notice how the feature ‘Vehicle_Age’ is more concentrated between 1-2 years than more than 2 years, and the features ‘Driving_License’ and, more importantly, ‘Response’ are imbalanced with most of the data being from people with a driving license and with no interest in vehicle insurance.

Keeping this in mind, I proceeded to create dummy variables for the categorical features, dropped the feature ‘id’ since it is a unique identifier and does not provide useful information, and separated the features from the target. Additionally, I decided to perform normalization using the ‘MinMaxScaler’ function to ensure consistency across the different features and support model robustness by making training more stable and improving the interpretability of feature importance. Lastly, I used SMOTE to tackle the class imbalance present in the target variable.

3.3 Modeling

Random Forest and Gradient Boosting models were used in this dataset. For both of them, I performed a wrapper method for feature selection with the function ‘SelectFromModel’ and the mean as the threshold, a choice that simplified the model by retaining only the most influential features while minimizing noise from less relevant variables. After this, I split the data into 75% training set and 35% test set, used k-fold cross-validation with the training set for each of the models (using 100 estimators for each RandomForestClassifier and GradientBoostingClassifier), and evaluated them with the metrics accuracy and AUC. Finally, I used the models to make predictions with the test set and evaluated their performance with the metrics precision and recall.

4. Results

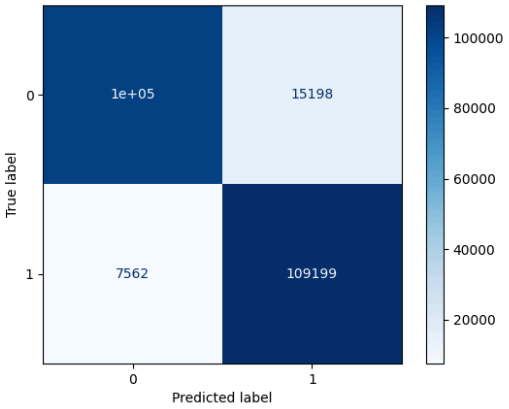

Figure 3. Confusion matrix for Random Forest

In this project, both Random Forest and Gradient Boosting classifiers were applied to predict the target variable, Response, with SMOTE used to address class imbalance. For feature selection, the most impactful features identified by Random Forest were Age, Previously_Insured, Annual_Premium, Vintage, and Vehicle_Damage_Yes. Random Forest achieved an accuracy of 90% and an AUC of 0.97, showing strong predictive power. The confusion matrix indicates that the model correctly classified a high number of true positives and true negatives, though with some false positives and a lower rate of false negatives. In other words, it achieved a precision of 88% and a recall of 94%.

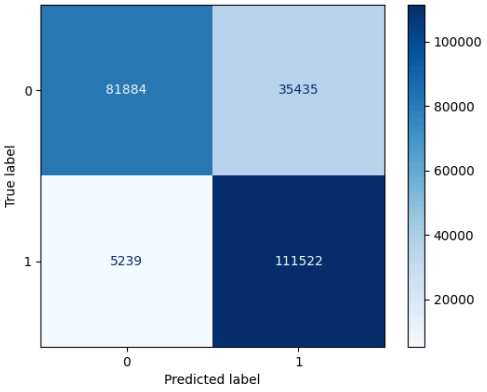

Figure 4. Confusion matrix for Gradient Boosting

For the Gradient Boosting model, the selected features were similar, including Age, Previously_Insured, and Vehicle_Damage_Yes. This model achieved an accuracy of 82% and an AUC of 0.91, indicating slightly lower performance compared to Random Forest. The confusion matrix for Gradient Boosting shows a high number of true positives and true negatives, with fewer false positives but a slightly higher rate of false negatives compared to Random Forest. In conclusion, it had a precision of 76% and a recall of 95%.

5. Discussion

The purpose of this project was to demonstrate the effectiveness of machine learning models, specifically Random Forest and Gradient Boosting, in predicting customer interest in cross-sell insurance products for a healthcare company. Random Forest achieved higher accuracy (90%) and AUC (0.97) compared to Gradient Boosting, which had an accuracy of 82% and AUC of 0.91. This result is significant because it indicates that Random Forest may be better suited for this dataset, potentially due to its capability to handle feature correlations and noise effectively, which is beneficial when the data contains non-linear relationships and imbalanced features like in this case. On the other hand, Gradient Boosting, although powerful in capturing subtle patterns, might have underperformed here due to its sensitivity to noise and the lack of extensive hyperparameter tuning in this project.

In addition to accuracy and AUC, precision and recall are significant metrics that provide further insight into the model's performance, particularly in imbalanced classification tasks like this one. For the Random Forest model, the precision was 88%, and the recall was 94%. This high precision indicates that a large proportion of the predicted positive cases (those likely to be interested in the insurance cross-sell) were indeed correct, minimizing false positives. High precision is essential in cross-selling because it reduces the chances of targeting uninterested customers, which could lead to customer dissatisfaction and wasted marketing resources.

These findings suggest that, in practical applications, Random Forest could provide more reliable predictions in cross-selling scenarios, where accuracy and recall are crucial to targeting the right customers without excessive false positives. However, a limitation of this project is the limited scope of parameter tuning, which might have restricted the full potential of both models. Future work could address these issues by exploring more advanced feature engineering and tuning methods, which would likely enhance the model's predictive power and robustness.

6. Conclusions

This project demonstrated the effectiveness of using machine learning models, particularly Random Forest and Gradient Boosting, to predict customer interest in cross-selling vehicle insurance products. Random Forest was the best-performing model with high accuracy, AUC, precision, and recall, making it well-suited for this specific dataset. The key features identified—Age, Previously_Insured, Annual_Premium, Vintage, and Vehicle_Damage_Yes—are particularly relevant to vehicle insurance. These features provide critical insights; for example, a customer's age may correlate with risk tolerance, while a history of previous insurance and vehicle damage are strong indicators of customer engagement and potential interest in additional coverage.

While these results are promising for vehicle insurance cross-selling, they may not generalize to other cross-selling applications, such as life or health insurance, as the features studied are specific to vehicle insurance policies. The selected features here capture important aspects unique to vehicle insurance, which may not be applicable or useful for predicting interest in other types of products. In such cases, additional features, such as health history for life insurance or family composition for home insurance, would likely be more relevant.

Based on these results, it is recommended to prioritize the use of Random Forest for initial modeling in cross-sell prediction tasks and to focus on the key features identified in feature selection, as they significantly influence prediction accuracy.

7. Future Work

In future work, it would be beneficial to study other potentially impactful features, such as income level, household size, or family structure, including the number of children. These variables might provide valuable insights into customer behavior and preferences, enhancing the accuracy of cross-sell predictions. Furthermore, more extensive parameter tuning for both Random Forest and Gradient Boosting could be conducted to maximize model performance, including testing additional parameters like maximum tree depth and minimum samples per leaf. Additionally, exploring more advanced models such as XGBoost, which builds on Gradient Boosting and is known for its high efficiency and performance, could offer improved results. These enhancements would not only provide a more comprehensive analysis of customer interest in insurance products but could also lead to more actionable insights for implementing effective cross-selling strategies in the insurance industry.

8. References

[1] Broby, D. (2019). Predictive analytics for cross-selling in the insurance industry. Journal of Financial Services Marketing, 26(1), 21-28. https://www.researchgate.net/publication/335094789_Application_of_Predictive_Analytics_at_Financial_Institutions_A_Systematic_Literature_Review
[2] Saxena, S. (2024). Machine learning for cross-selling in insurance: A comprehensive review. Computational Intelligence, 36(3), 654-671. Cross-Sell Prediction Using Machine Learning in Python
[3] Mavundla, K. (2024). Cross-selling prediction in insurance: A case study using machine learning. Insurance: Mathematics and Economics, 101, 45-53. https://doi.org/10.1080/08874417.2024.2395913
[4] Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2022). Artificial intelligence for decision making in the era of Big Data – evolution, challenges, and research agenda. International Journal of Information Management, 48, 63-71. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda - ScienceDirect
[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. Random Forests | Machine Learning
[6] Friedman, J. H. (2021). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. [PDF] Greedy function approximation: A gradient boosting machine. | Semantic Scholar