Health Insurance Cross Sell Prediction

Samuel Sears
Summer 2021
https://github.com/ssears219

Domain

This data is coming from the insurance domain. Here are some references to make sense of what I am planning to do with the data.

1. https://risk.lexisnexis.com/insights-resources/case-study/maximize-property-to-auto-cross-sell-campaign-performance - Property to Auto insurance cross sell case study by LexisNexis
2. https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction/discussion - Discussion board with 12 posts clearing up any confusion about the task and data set
3. https://gutentagworld.wordpress.com/2020/12/13/health-insurance-cross-sell-prediction/ - Project using the same data set. I will be able to use this to improve on the results.
4. https://support.sas.com/resources/papers/proceedings17/0941-2017.pdf - Research paper providing a step-by-step approach to maximizing cross-sell opportunities with financial products.
5. https://www.ibm.com/blogs/nordic-msp/six-ways-use-data-science-drive-cross-sell-upsell-activity/ - Blog post from IBM on ways to drive cross-sell and upsell activity.
6. https://blog.aspiresys.com/digital/big-data-analytics/boost-your-sales-by-cross-selling-and-up-selling-using-predictive-analytics/ - Blog post from Aspire Systems on cross-selling and upselling using predictive analytics.
7. https://www.psmbrokerage.com/blog/7-tips-on-how-to-cross-sell-insurance - Blog post on cross-selling insurance geared towards insurance agents. This will give me insight into how an insurance agent may approach this task outside of predictive analytics.
8. https://www.agencybloc.com/resources/grow-your-agency/selling/how-to-cross-sell/ - Another blog post on cross-selling insurance that is geared towards insurance agents.
9. https://www.kaggle.com/dmkravtsov/11-0-health-insurance-auc-0-97 - Submission that used XGBoost to obtain 0.97 AUC.
10. https://www.kaggle.com/yashvi/vehicle-insurance-eda-and-boosting-models - Another submission using XGBoost, LGBM, and CatBoost classifiers.

Data

Link to Kaggle Dataset - https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

The dataset includes test, train, and sample submission CSV files. The data is composed of a unique id column, 10 predictor variables, and one response variable. There are 381,109 train records and 127,037 test records for final submission. For this project, I will only be using the train dataset which has the response variable included.

Variable	Type	Description
id	Int	unique identifier
Gender	Categorical (Male, Female)	Gender of customer
Age	Int	Age of customer

Driving_License	Boolean (0, 1)	0 – Customer does not have a driver's license 1 – Customer has a driver's license
Region_Code	Int	Unique code for region of customer
Previously_Insured	Boolean (0, 1)	0 – Customer doesn't have vehicle insurance 1 – Customer does have vehicle insurance
Vehicle_Age	Int	Age of vehicle
Vehicle_Damage	Boolean (0, 1)	0 – Customer didn't get vehicle damaged in past 1 – Customer did have vehicle damaged in past
Annual_Premium	Int	Annual premium customer pays for health insurance
Policy_Sales_Channel	Int	Anonymized code for channel of outreach to customer
Vintage	Int	Number of days customer has been associated with insurance company
Response	Boolean (0, 1)	0 – Customer is interested 1 – Customer is not interested

Research Questions and Benefits

This project will simulate a real-world problem where an insurance company is attempting to identify cross-sell opportunities within it's current customer base. Specifically, the goal will be to identify health insurance customers who are interested in vehicle insurance. Based on the dataset provided, the health insurance company has already reached out to customer in the past via different sales channels (Policy_Sales_Channel) and tracked whether the customer was interested in vehicle insurance or not (Response). Being able to predict whether or not the customer will be interested given the factors in the dataset will help the company optimize their communication and sales process to optimize its business model and maximize revenue.

Methods

There are two main tasks I will be completing with this project. The first task will be to explore the data and find any customer segments that have a high rate of interest. This descriptive analysis will provide value to the business in understanding their customer base. Next, I will use the insights I gained from the descriptive analysis to build a binary classification model that predicts the Response variable. For the model, I will test a variety of models including Naïve Bayes, Logistic Regression, Decision Tree, XGBoost, and a Neural Network. The main evaluation metric I will use for the models will be AUC.

Potential Issues

I plan to use python so there will always be package installation issues I will have to work through. However, I am confident I will be able to remediate those issues or find other paths forward.

This is a highly unbalanced dataset. Unsurprisingly, over 87% of the data is of the not interested in vehicle insurance class. Furthermore, it will be more important for the model to identify the minority class. With such class imbalance, the model will have a hard time learning characteristics of the minority class and performance will suffer. I plan on utilizing the SMOTE, Synthetic Minority Over-sampling Technique, to solve for this problem.

The final risk to highlight is the abundance of work. It will be important to leave time for predictive modeling and not get too carried away with the descriptive analytics task. Time will be allotted for potential package installation and environment management troubleshooting, training, and model optimization.

Conclusion

In the insurance domain, writing new policies equates to increased revenue. To sell a new policy, the company can obtain a new customer or sell a different insurance product to an existing customer. Given the company typically already has data available to leverage for an existing customer, there is a major opportunity to be more targeted in pursuing sales of additional insurance products to an existing customer.

In this project, I will simulate a data science project in which an insurance company is attempting to identify current health insurance customers who would be interested in vehicle insurance. I will analyze past data of customers who were recipients of a cross-sell in order to inform the business on the customer segments with the most opportunity. Furthermore, I will develop a model that will predict whether or not the customer would be interested in vehicle insurance depending on the variables that would be known prior to solicitation.