Insurance Cross-Selling

With Predictive Analytics

Team: ggplot4

BACKGROUND & MOTIVATION

For insurance companies, sales can be one of the most part of their businesses. More policies sold does not only mean higher revenue, but also lower expected risks for the insurance companies when unusually large claims are reported. However, new lead generation could be labor-intensive and ineffective. Therefore, insurance companies are always looking for ways to cross-sell insurance products to their existing customer base. Knowing who to reach out to and who to avoid will allow for the higher sales numbers as well as avoiding wasted marketing efforts. Using existing customer information for sales leads will help to drastically lower customer acquisition costs for the company. This frees up resources and capital for other projects to acquire even more customers.

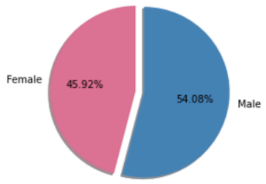
This report looks at how a predictive analytics model can impact insurance cross-selling results and increase revenue. Using real insurance customer data, a predictive model can be created to see who will be more likely to sign up for an auto policy in addition to their current health coverage. Finding better sales leads from within is crucial to an insurance company looking to increase their cross-selling results.

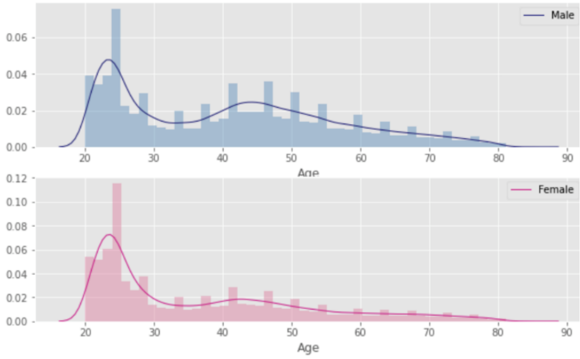
DATA EXPLORATION

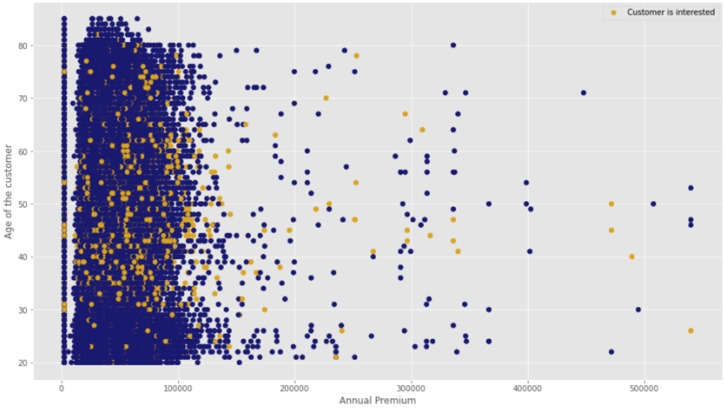
We used a dataset from an insurance company that was available on Kaggle. It includes consumer features about demographic information (age, gender, region, driver's license status), vehicle-related information (vehicle age, past vehicle damage status) and insurance policy information (previous insurance status, current annual premium, policy sales channel, customers' tenure with current carrier).

First, by doing an Exploratory Data Analysis, we are trying to get a big picture of how existing consumers look and make some sense of it.

Of the 762,218 customers in the dataset, the sex ratio is nearly 1:1. 46% of current customers are female, while 54% are male. Looking at the data by age, the histogram has a right-skewed distribution and a long tail, showing that customers are highly skewed towards younger generations, and most customers are in their twenties.

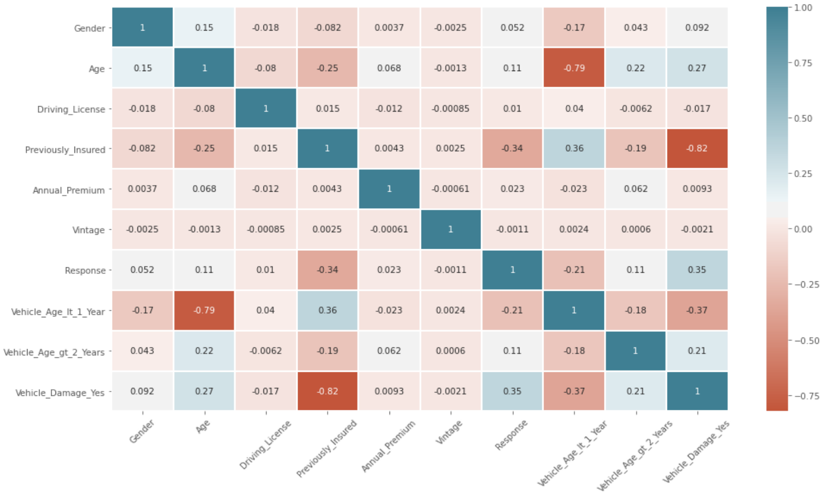




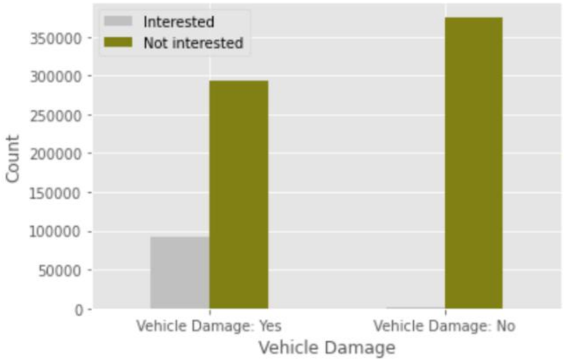


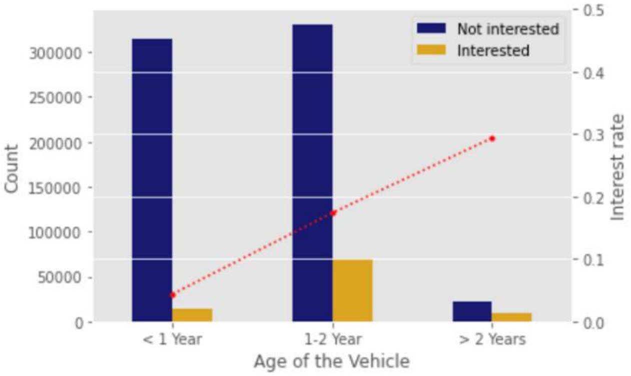
We also investigated in possible relationship between age, the amount paid by the customer as premium for the year, as well as their interest in vehicle insurance. However, when looking at the scatter plot, we see that there is no significant connection between these variables. The amount paid by the customer as a premium for the year is typically less than 100,000, not clearly associated with customer age or their interests in vehicle insurance.

Looking at the correlation heatmap, the relationships between features become clearer. Past vehicle damage status positively correlated to the target variable response, indicating that customers who got his/her vehicle damaged previously are more interested in vehicle insurance. On the other hand, previous insurance status and vehicle age are negatively correlated to the target variable response, indicating that customers who already have vehicle insurance and whose vehicle age < 1 year is less interested in vehicle insurance.



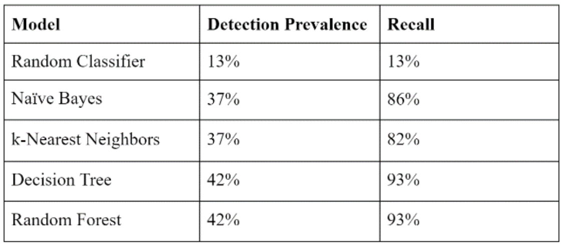
Looking at the key features in more detail, we found that consumers who have suffered vehicle damage in the past are 50 times more likely to be interested in vehicle insurance than those who have not suffered vehicle damage in the past. The interest rate is also linearly associated with the age of the vehicle. The older the vehicles are, the higher their interest in vehicle insurance. With a basic idea of how the dataset looks like in mind, we started the process of building and evaluating models.





OUR MODEL

Prior to building the models, we manually selected features based on the exploratory results and the attributes' interpretability. For example, we excluded the policy sale channel attribute since it is an anonymized attribute, and we could not draw meaningful and widely applicable conclusions from it. After selecting features to keep, we under sampled the training dataset to prevent models from having bias towards the majority class. We compared several models to compare the percentage of customers each of the classify as interested (detection prevalence) as well as the percentage of actual interested customers each model captured (recall).

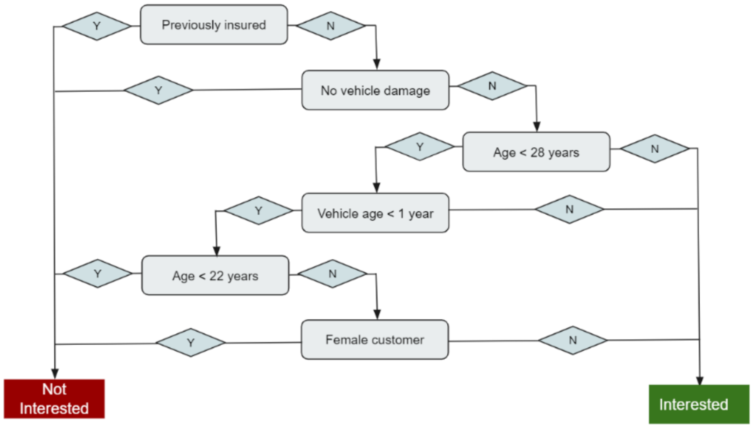


Model	Detection Prevalence	Recall
Random Classifier	13%	13%
Naïve Bayes	37%	86%
k-Nearest Neighbors	37%	82%
Decision Tree	42%	93%
Random Forest	42%	93%

We decided that a decision tree model most effectively predicts interested prospects to whom the company should engage with cross-selling efforts. By predicting less than half of the overall customer base as interested customers, it can capture 93% of truly interested customers with less complexity than that of Random Forests'. Both Naïve Bayes and k-NN have lower recall without significantly lower detection prevalence. Besides comparably better performance, the visual nature of the decision tree model output also allows us to give comprehensive recommendations.

The model predicts three types of customers as being interested in auto insurance:

- Customers who are older 28 or older who have not had vehicle insurance before and have had prior damage to their vehicle
- Customers who are under 28 years of age who have not had vehicle insurance before and have had prior damage to their vehicle that is more than one year old
- Male customers between 22 and 28 who have not had vehicle insurance before and have had prior damage to their vehicle that is less than one year old

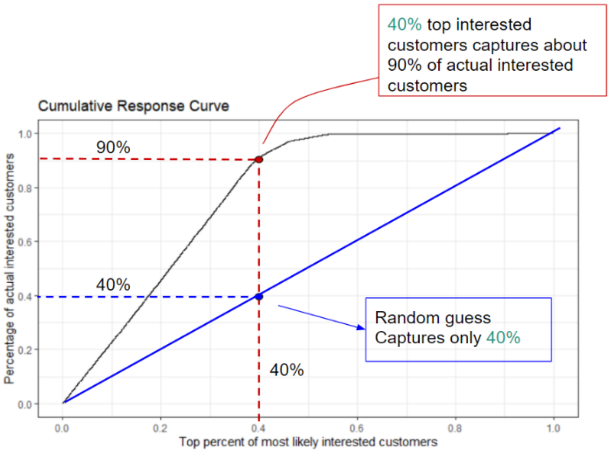


RECOMMENDATIONS

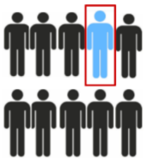
Now that the model has been built it can now be put into use by the business. Business leaders in sales and marketing can use this model to help target customers that are most likely to sign up for the vehicle insurance. This means there can be a higher return on advertising and sales campaigns. Higher success rates for the sales team will allow them to use their time more wisely by acquiring even new customers or additional bundles for their current customers. Lowering customer acquisition costs is also a major impact from using this model. Leveraging existing customer data will help to increase revenue and stay ahead of competitors. Using this predictive analytics approach there will be real dollar value created by the business that far exceeds a random approach.

COST BENEFIT ANALYSIS

Adopting recommendations derived from the model could improve profits by almost $2 million if not more. Our model can find 90% of the true interested customers by only looking at 40% of the customer base who do not already have auto insurance, without sending a survey. To illustrate using a hypothetical example, if this company has 2 million customers in this group, based on past survey data, we can expect 12% of them to be interested. We are assuming interested customers are more likely to buy auto insurance than uninterested, let's say interested customers buy auto insurance 5% of the time, instead of 1% if they are uninterested. This is like selling to a warm lead. Conversion rates go up. When we are dealing with big numbers, small differences make a big difference. In this case, our model finds 216,000 of the true interested customers. Of these, about 10,800 will buy auto insurance. If it costs $3 per targeted sale and revenue equals $1500 per new auto policy, then profit here would equal about $13.8 million, whereas targeting everyone would yield a profit of $12 million. Randomly selecting 40% would yield a profit of $4.8 million.


40% top interested customers captures about 90% of actual interested customers

Random guess Captures only 40%



Appendix

R Markdown for Predictive Model

Jupyter Notebook for Exploratory Data Analysis