Optimal Premium model

Context

- **This preliminary analysis** was made to show to an Insurance Company the value of building an effective model for sales, using past transactional data and socioeconomic data of several companies obtained from surveys.
- The fitted model will be able to identify the **customers with most propensity to make a purchase.**

Objectives

- Present the descriptive analysis of the data.
- Describe the models used to solve the problem.
- Show the value of **the first fitted models** and further improvements.

Data preparation: scope & cleansing process

Target

- Develop a predictive model that using past data allow us to determine which customers has more probability of make a purchase after a call. Then, use these results to determine the total premium obtained for each client.

Datasets and Variables to use

- One dataset with: data obtained from online surveys.
- One dataset with the target variable: Sales, Premium offered and Past transactional data.

Cleansing process

- The acquired data has been prepared in the following way:

1. **Transform some variables in categorical/binary data** (ex. state, industry, subindustry, business structure, between others) **or numeric data** (year established).
2. **Filtering and imputation**: inspect the quality of the variables. For example, the variable region was modified to maintain the code description values. Also, we retained only the sales where we have confidence of a win/loss sale. NA's were less than 1%.
3. **Cleaning inconsistencies:** Some variables were not considered in the model due to inconsistencies, or due redundancy (ex. industry vs. subindustry), or due to poor predictive power or high correlations values (multicollinearity/noise).
4. **Creating new variables:** new variables were created and we include interactions between couple of variables that can help to find more complex hidden patterns in the data that may improve the results.

Correlations

Here we show the **correlations of a selection** of variables:

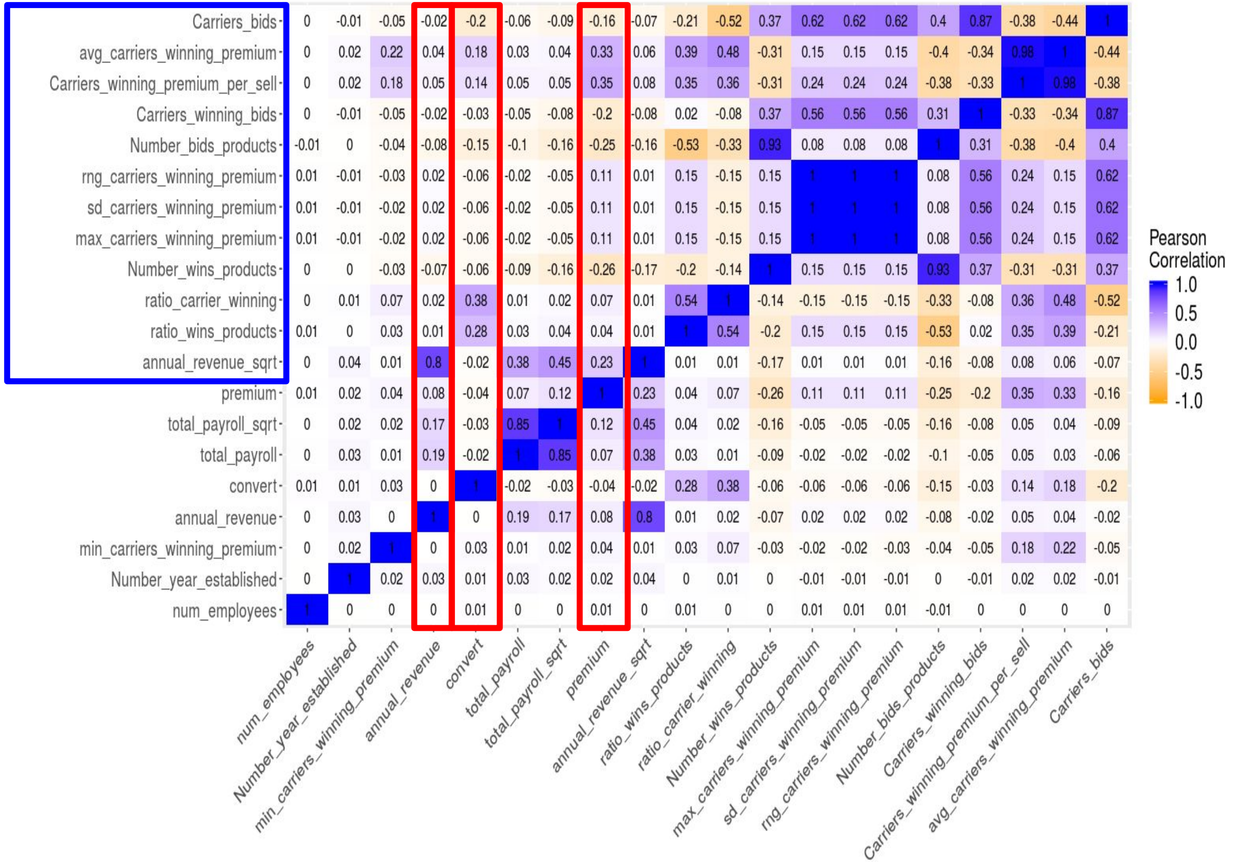
- The variables more related with **convert** were created using other variables: carrier's ratio of winning bids and product's ratio of winning bids

- The variables **more related** with **Premium** are also variables created using other variables. Ex. The premium offer for the winning carriers.

- We use different transformations in order to get better correlations for total payroll and annual revenue.

- **Numerical variables** obtained from the surveys seems not show strong correlations with the variable to predict (**convert**)

New variables (some of them):



Dataset: Featuring Engineering

Actual Solution:

Predict Sales Outcome

1. Predict Sales outcome (**1: Win a sale**, **0: Loss a Sale**) for a specific Premium.
   - Increase sales information using the following rule:
     - If you gain a sale with a price: **p_high**, this implies a win at all lower prices.
     - If you lose a sale with a price: **p_low**, this implies a loss at all higher prices.

This adds win/loss sells in approximately the same ratio as the original data.

Then we calculate the **account value** as the sum of the premium by the convert value (value between 0 and 1).

Possible improvement:

Optimize the offering premium

1. Introduce into the model above all the possible premiums**,** this will give the probability of win the offer.

2. The **optimal price** to get the **maximum profit** can be determined by **maximizing expected profit given the underlying seller costs**. In this case, the seller costs are supposed to be zero. So, our optimization problem would be maximize the following function:

$$Profit(X) = P(Sales = 1|X)[Premium - Cost]$$

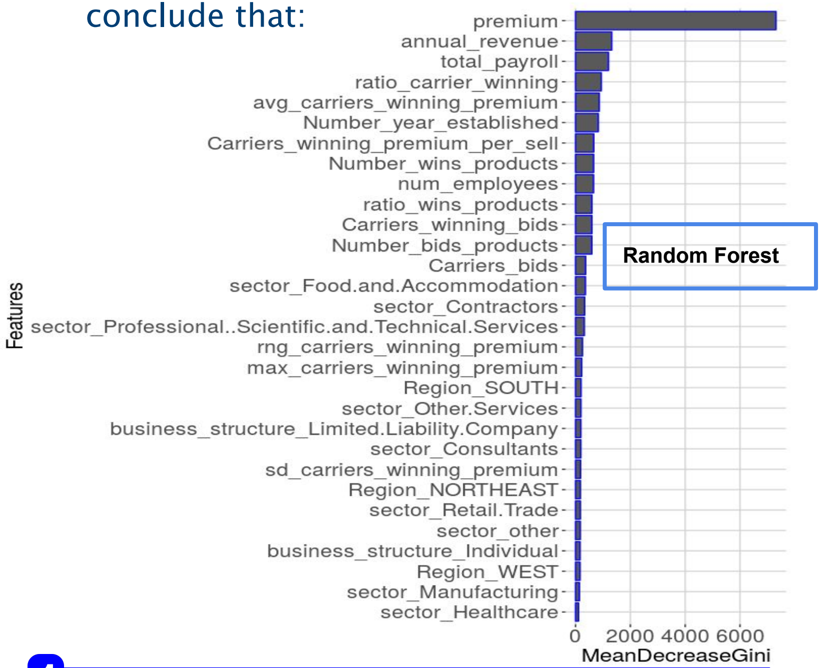
Applied methodology

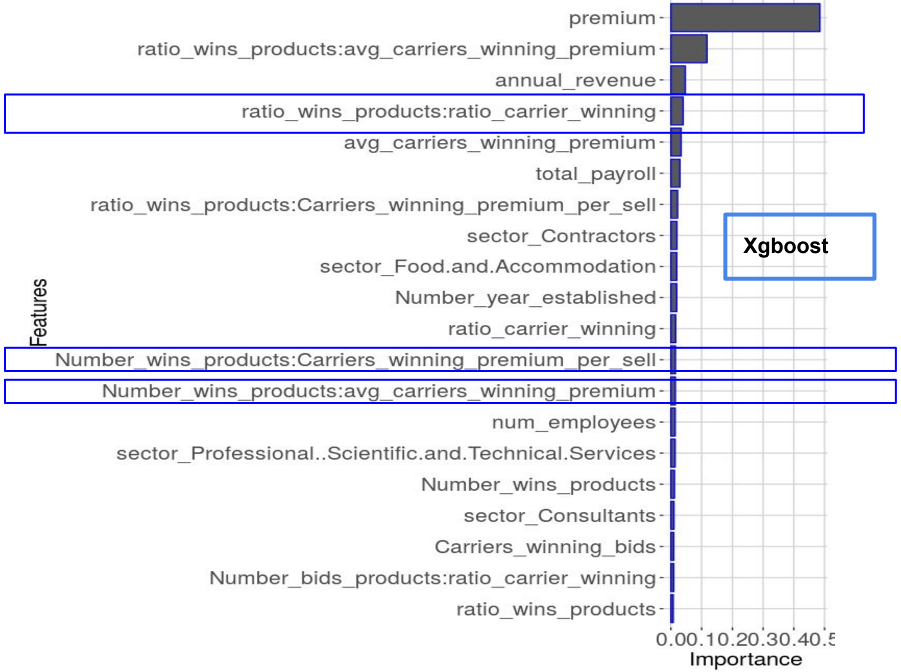
The process to create a powerful and robust predictive model relies on **the following steps:**

Step	Description	Result
Handling imbalanced data	The data does not seems severely imbalanced.	No Sales aprox. 42% Sales aprox. 58%
Algorithms comparison & selection	We used **Boosted decision trees (XGB)** and **Random forest (RF)** to predict the probability of a successful sale.	Value to minimize: **AUC**
Parameters optimization	We spent some time to optimize the parameters of the model **Boosted tree** in order to obtain the best results.	**10-fold cross val** XGB: 0.833 RF: 0.824
Robust validation	First, we used **5-fold cross-validation** technique to train and test the models. Later, we **tested the model again** on an independent test set, not used for the training.	**AUC** XGB: 0.70 RF: 0.74

Dataset: Featuring Selection

- We use two methodologies **Random Forest** and **Xgboost.** Looking the importance of the variables we can conclude that:





1 We can see that premium, annual revenue, total payroll seems to be the most important variables for the model. New variables like ratio of winning of a carrier show to be also important.

2 We investigate some **interactions** between couple of variables. This can help to find more complex hidden patterns in the data that may improve the results.

Results

With a **test set of 10%** of the original data, the model seems to **validate the past pattern of the sales** (confusion matrix). Now, with the **second dataset** we will **estimate the value** of the model based in the performance of the theoretical model in the future (expected theoretical value).

Test Set (10%)

Confusion matrix (0.52)*

***Optimal threshold**

Random Forest (RF)

		Prediction		
		Loss	Win	**TOTAL**
Actual Value	Loss	309	183	492
	Win	194	485	679
TOTAL		503	668	1171

Accuracy	**0.6780**
Precision	**0.7260**
Recall	**0.7142**
f1-score	**0.7201**

Boosted Tree (XGB)

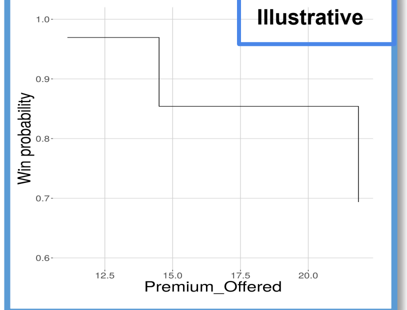
		Prediction		
		Loss	Win	**TOTAL**
Actual Value	Loss	268	224	492
	Win	193	486	679
TOTAL		503	668	1171

Accuracy	**0.6438**
Precision	**0.6845**
Recall	**0.7157**
f1-score	**0.7000**

Profit

When you give to the model all the possible premiums, the model tends to give more probability to the lowest one

To choose the **optimal price** we will look to the **expected profit value** instead the probability of win a sale (although several scenarios may be created).



- Another way to compare the quality of our results is compare them with a baseline model, where a sale is the result of the toss of a coin. With this random model we can validate if there exist an improvement with the model in comparison with doing nothing (Further improvement).
- **The best RF results seems to better ones.**

Conclusions & Recommendations

- The results seem indicate that the model with **best performance is the RF**. In order to maximize the TP and minimize the FP we consider a threshold of .52 for the results. Improvements over this model and implementations of new ones, including ensemble models are further improvements.

- The **performance of the model could be improved** by incorporating additional features which include dynamic interactions (time series), and more interactions between variables, also a more complete data of the companies, price sensitivities and a quantification of price-elasticity curves

- In the future, it might be possible to **perform a controlled experiment** using the results of the machine learning algorithms and socioeconomic data, to generate more cases for the premium that allows find the optimal profit based on past data.