Predicting Caravan Insurance: A Bumpy Ride

Fabian Blasch, Gregor Steiner, Sophie Steininger, Jakob Zellmann

02/28/2022

1 Introduction

From the perspective of a marketing analyst, it would be great to know which customers are interested the most in obtaining an additional product to specifically focus marketing activities on these customers. This premise is the focus of our analysis on Caravan Insurance Prediction. The dataset used was obtained from Kaggle, a website for data scientist which hosts various data science competitions and also published datasets for educational purposes. The Kaggle Competition in focus here is the CoiL Challenge on Caravan Insurance Prediction. It focuses on predicting whether already existing customer of an insurance firm would be likely to obtain a caravan insurance policy. This prediction would be used to determine which customers would receive marketing mail for caravan insurance. Therefore, the main question to be answered during the competition by the provided data is: "Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?" This question consists of two parts, the predicting problem and identifying the features with the largest explaining power.

For the evaluation of the winner great emphasis was put on having as little false negatives as possible in the predictions. This focus was due to the characteristic of the problem as not sending a mail to a customer, who would purchase an insurance policy, is much costlier than sending out a mail to customer that is not interested in purchasing insurance. There are also no other negative effects associated with sending the mail than the cost itself. Another reason for this focus in evaluating the winner was that the data set is highly unbalanced, as only a small number of customers is interested in caravan insurance. This would lead to a correct prediction of around 94%, if you simply assume no one is interested in this kind of insurance. (Elkan 2001) As this approach does not lead to any interesting insights from a marketing perspective, the primary focus was on minimizing false negatives in the predictions. Accordingly, for performance evaluation, competitors were asked to send in 800 observations carrying the highest probability of purchasing caravan insurance when contacted. The true positives of those 800 observations were then interpreted as the score of the competition.

During the competition the data set was originally split up into two parts. During the competition only the training set was released to the public and the evaluation set was used to determine the winners. Now, after the competition is finished, both sets are combined in the publicly available Dataset on Kaggle, which allows us to compare our predictions to the competition.

The rest of the paper is structured as follows: Firstly, the dataset used is described in detail, then the methodological approach is explained and then the corresponding models are presented. Finally, we are going to conclude with our best predictions.

2 Data description

The data used in this Kaggle Competition is owned by the Dutch datamining company Sentient Machine Research and is based on real world business data. The dataset includes 86 variables on product usage and socio-demographic features. Each variable represents the value within a postal code area. The training data set includes over 5000 descriptions of customers while the evaluation data set contains around 4000 values.

The socio-demographic variables allow for differentiation between the differences in the postal code areas. These variables include variables such as average age or number of houses. Additionally, there are numerous features on socio-demographic variables that represent a percentage within each group such as percentage of religions and marital status. An important note here is that the socio-demographic variables are taken from the average in the postal code areas. Therefore, the customers are matched to the area they live in, and the values assigned are the averages from that region. The advantage of this approach is that the socio-demographic variables can easily be matched to the existing customer database but does not contain as much information as individual data would have.

The product usage features contain two types of information, firstly absolute contribution to policies and secondly number of policies held. From a theoretical perspective, these two kinds of features are collinear as customers with a higher number of policies in a specific insurance field are also having a higher absolute contribution. However, they cannot be combined as the information on the premium paid per policy can only be identified when both features are included in the models.

Additionally, the data was already discretionized, meaning that all numerical values were transformed into discrete variables. There were four groups introduced: customer subtypes, average age key, customer main type and keys for percentages.This grouping into levels is usually performed to allow for easier interpretation and increasing the correlation of the feature with the dependent variable, as the number of freedoms is infinite in continuous data. Additionally, many prediction models do not perform well with continuous data, for example random forests. (Gupta 2019)

Another characteristic of our data set is the unbalanced dependent variable. Of the 5822 customers in the training data only 348 owned a caravan insurance, which is only 6%. Therefore, there is too little positive data for some predictive methods, especially learning algorithms. We attempt to overcome this problem by over- and under-sampling in a later section.

Overall, the Data is noisy, skewed, collinear and high dimensional with a weak relationship between independent and dependent variables. Additionally, one could expect that the socio-demographic data used is not going to be as informative as the behavioral data on insurance policies. The socio-demographic features used are derived from the postal area and do not contain information on an individual level, while the data on insurance policy contains information on individual behavior. Therefore, in comparison it would be reasonable to expect that the product usage features outperform the socio-demographic ones. (Elkan 2001)

3 Models

3.1 Elastic Net GLMs

The first type of model we use is a generalized linear model (GLM) with Elastic Net penalization. Since this model class was not explicitly covered in the lecture, we will give a brief introduction here.

The task is binary classification, therefore our dependent variable $y_i \in \{0, 1\}$ for all individuals $i = 1, \dots, n$. Furthermore,

$$y_i \sim Bin(1, p),$$

where $p = \mathbb{P}(y_i = 1) = \mathbb{E}[y_i]$. Thus, the model can be formulated as

$$\mathbb{E}[y_i] = g(X_i\beta),$$

where X_i is the i-th row of the design matrix X, β is the vector of parameters and g is the link function, which maps the linear predictor to $[0, 1]$. There are different options for the link function g. We try 4 different ones: Normal cdf (“Probit”), logistic cdf (“Logit”), Cauchy cdf and the complementary log-log function (“Cloglog”). Figure 1 provides a good illustration of how the different link functions map the linear predictor to the $[0, 1]$ interval.

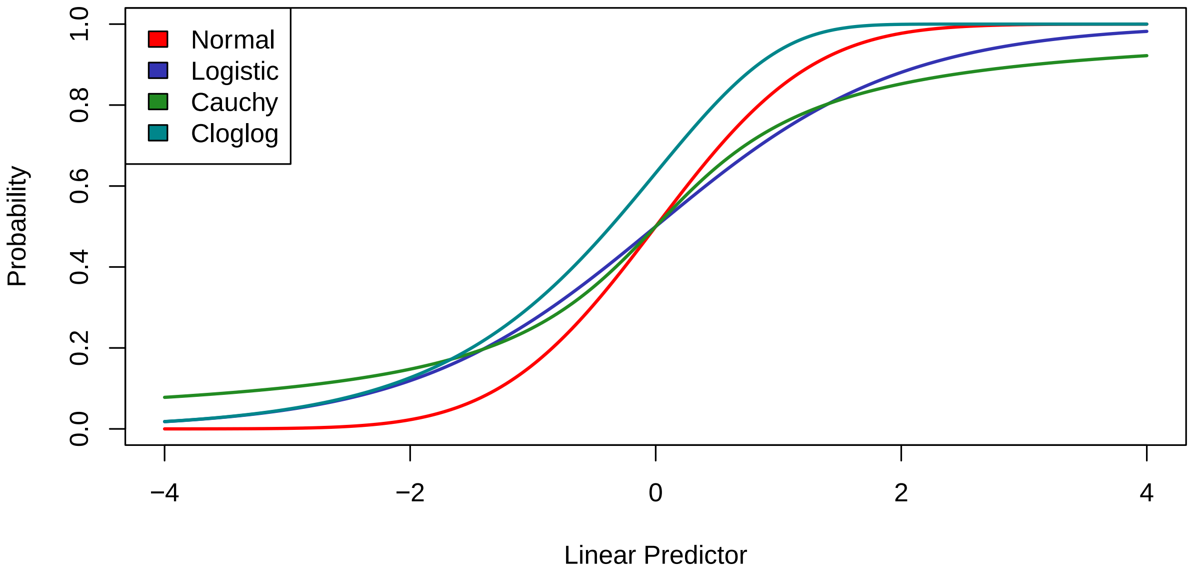

Figure 1: Link functions for the Binomial GLM

Furthermore, we use Elastic Net penalization. That means the penalty term is a convex combination of the Ridge and Lasso penalty terms. More formally,

$$C(\alpha, \beta) = \alpha||\beta||_1 + \frac{1-\alpha}{2}||\beta||_2^2.$$

This combines desirable properties of both Lasso and Ridge penalization. For example, due to the geometry of the Lasso, coefficients are shrinked to zero, which allows for explicit model selection, and Ridge provides performance improvements over unpenalized estimators for at least some values of the penalization coefficient λ and does especially well with highly correlated features.

The optimization problem to solve is

$$\min_{\beta \in \mathbb{R}^k} -\frac{1}{n} l(y, X; \beta) + \lambda C(\alpha; \beta)$$

where $l(y, X; \beta)$ is the log-likelihood. For more details see Friedman, Hastie, and Tibshirani (2010).

3.2 Tree based models

Furthermore, we will try different tree based models, in particular Random Forests and Extreme Gradient Boosting (XGB). These two methods were extensively discussed in class, therefore we will not give an introduction here.

4 Results

This section will present the main results of our project. We will begin by presenting the base models, i.e the models estimated using the raw unchanged input data, then we will discuss an approach for feature engineering and compare the results obtained with the original data to slight modified data. Subsequently the results of the different models applied in the frame of our analysis will be compared against each other and in reference to the Kaggle competition to the winning model as well. Before closing this section with a stylized model that allows to target customers optimally we will briefly discuss which features have the highest descriptive/predictive power.

4.1 Base Model

To compare the models across different cut-off values, the model selection process as well as the performance assessment will be based on Precision-Sensitivity and RoC Curves. First off, one may be interested in the difference in performance without an attempt to correct the imbalance of the data and also prior to any feature engineering. The following plots allow for a comparison of the models mentioned in the methods section, across a range of cut-off points.

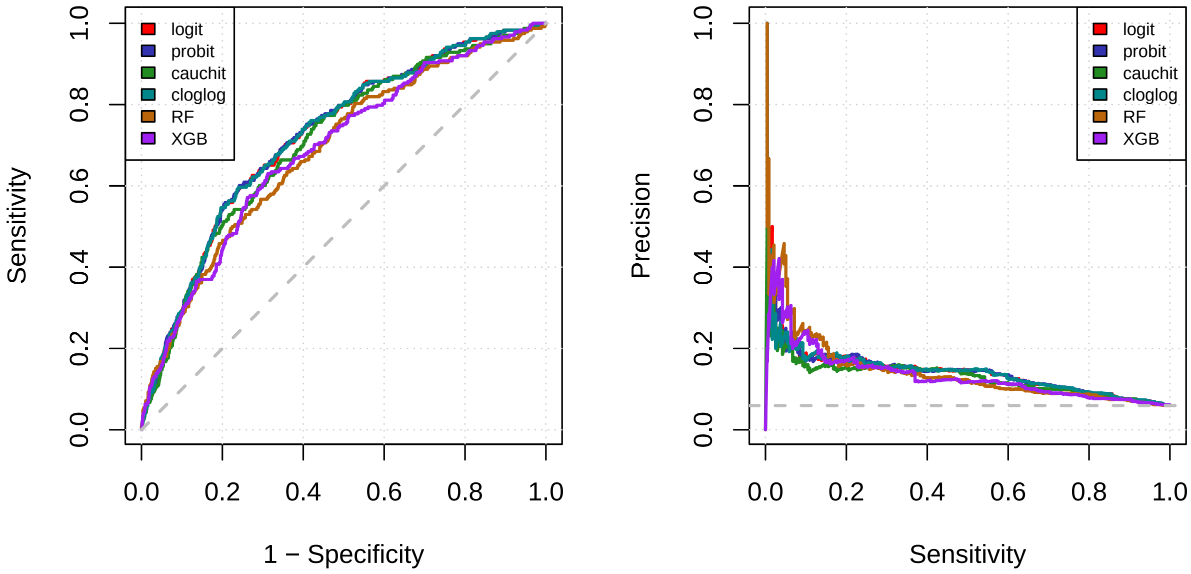

Figure 2: Model Comparison

In terms of the Sensitivity and Specificity trade-off one can observe that the Glmnets offer superior performance for virtually all cut-offs. Further, when comparing within the model class, the cloglog and probit link function seem to outperform the remaining link functions. Since the data set is quite imbalanced, i.e. only around 5.95% of the observations end up taking out caravan insurance, the trade-off between the precision and sensitivity is of even greater importance in our case. The rationale behind this is that precision and recall are not affected by the number of true negatives. When taking a closer look at the PS curve, we observe that the Glments no longer dominate the remaining models for all cut-offs, however, upon careful consideration of the background of this challenge we may still conclude that the Glmnets offer superior predictive performance. The PS curve clearly indicates that tree based models offer a better trade-off between sensitivity and precision than the Glmnets for high cut-off values.However, in the context of offering people insurance, the cost associated with the offer is usually drastically lower than the opportunity cost of not offering someone insurance, who would have purchased it. Accordingly, within the scope of this competition, we are willing to decrease the precision for higher sensitivity, or put differently, we want to increase the amount of true positives, knowing that we will simultaneously also suffer a higher number of false positives. Thus, even though the Glmnets do no longer offer superior performance across all cut-offs, the performance for the most relevant ones is still greater when compared to the other models. We may therefore conclude that the Glmnet utilizing the complementary-log-log link offers the best performance prior to an attempt to correct the imbalance in the data.

4.2 Feature Engineering

As mention above, the data solely contains categorical variables. This comes with some challenges for feature engineering as new feature can (to our knowledge) only be generated by creating interaction terms or reducing the levels of a feature. As the encoding of the feature is unknown, only the first approach remains relevant. Following the competitions winner (Elkan 2001) we construct two addition feature (an interaction term for the number of policies and the contribution per policy for car and fire insurances) and dismiss the socio demographic variables contained.

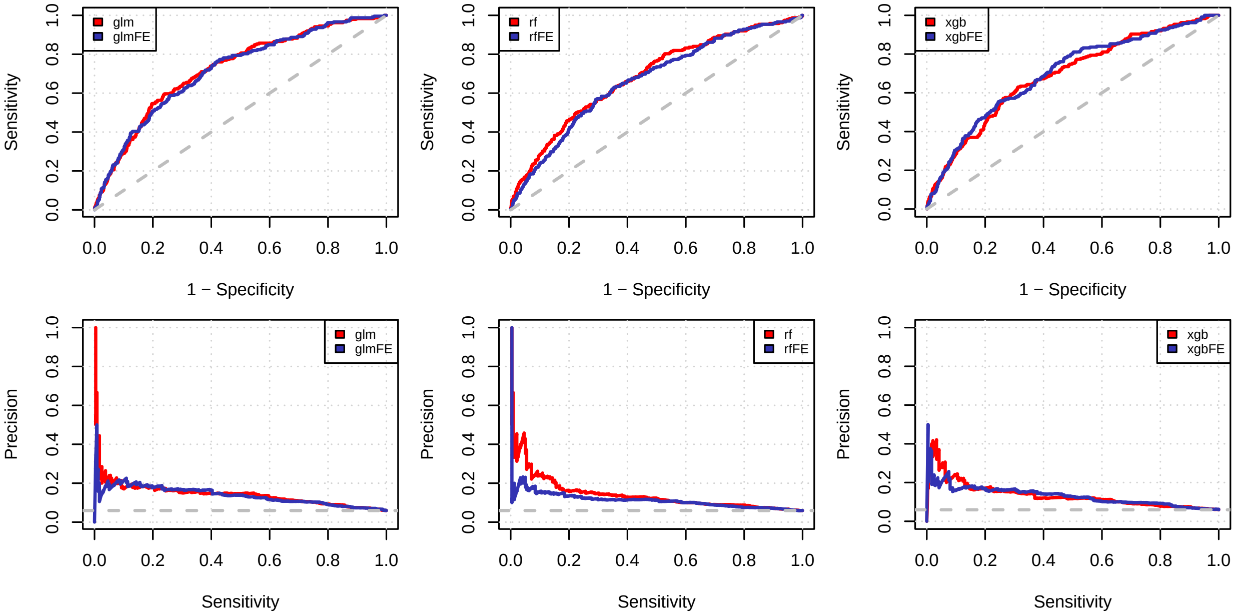

Figure 3: Feature Engineering

Above in the first row we see that except for the XGB model the RoC curve for the models are better with unchanged data (red). As the data is heavily imbalanced the second row, containing the Precision-Recall curves, gives more important information about the quality of the models. Here we find that for low cut off points, clearly all models perform better with the original data.

We therefore conclude that, at least in the set up of our models, the approach proposed by (Elkan 2001) does not lead to better results. As we could not come up with better ideas for feature engineering we therefore continue the analysis on the basis of the unchanged data.

4.3 Correcting the Imbalance

One way to tackle the imbalance in a data set is to synthetically re-balance the data via sampling methods. This can be achieved in a multitude of ways, however, we will only distinguish between over-sampling, under-sampling and a combination of the two.

Over-sampling means sampling from the minority class with replacement while leaving the majority class intact. Under-sampling refers to a sampling process in which one samples from the majority class without replacement. The remaining option is to combine over- and under-sampling.

4.3.1 Glmnets

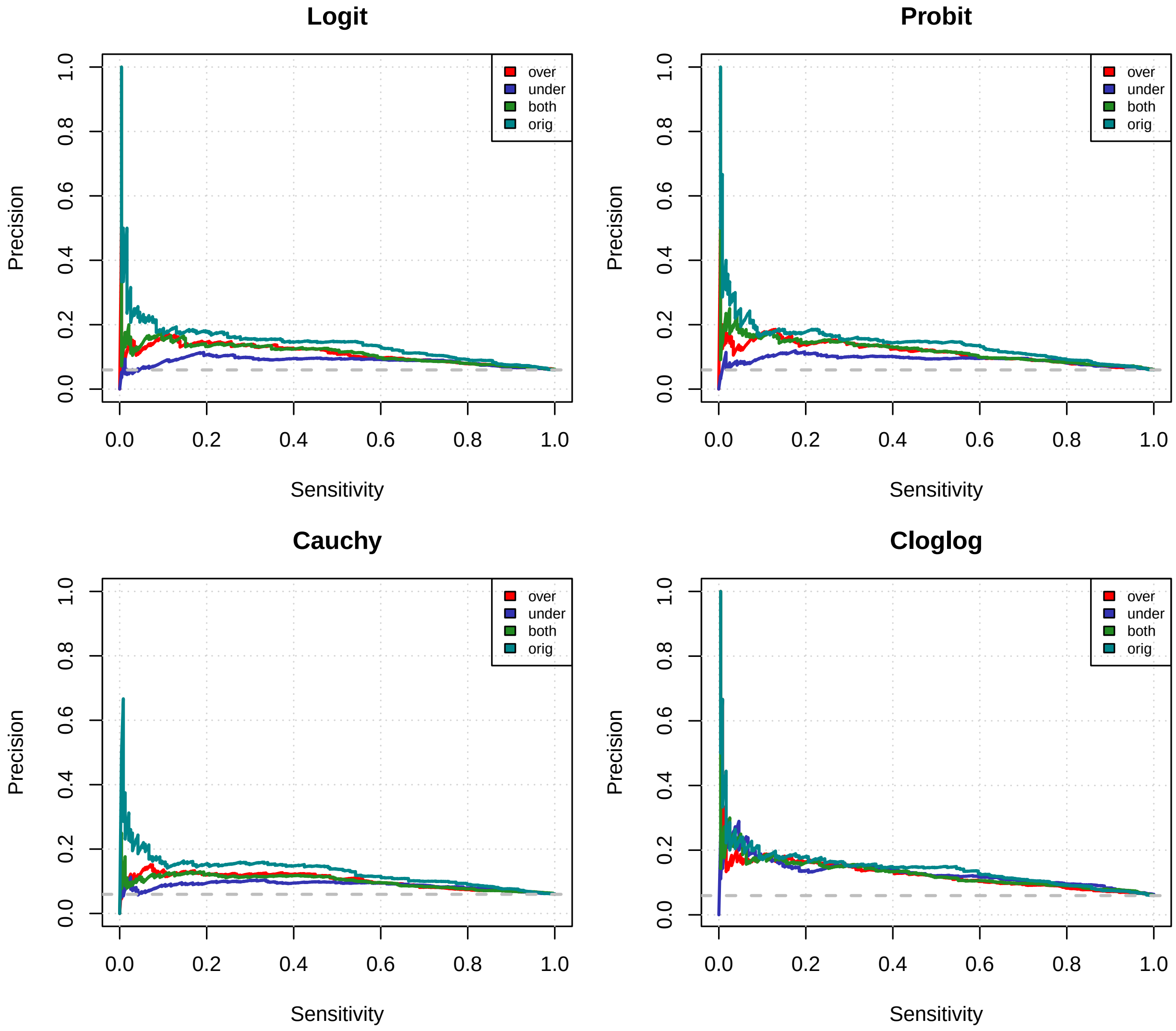

Figure 4: Glmnets resampled

The PS-curves above clearly depict that the balancing via re-sampling did not have a positive impact on the trade-off between precision and sensitivity, independent of the link function.

4.3.2 Random Forest and XGB

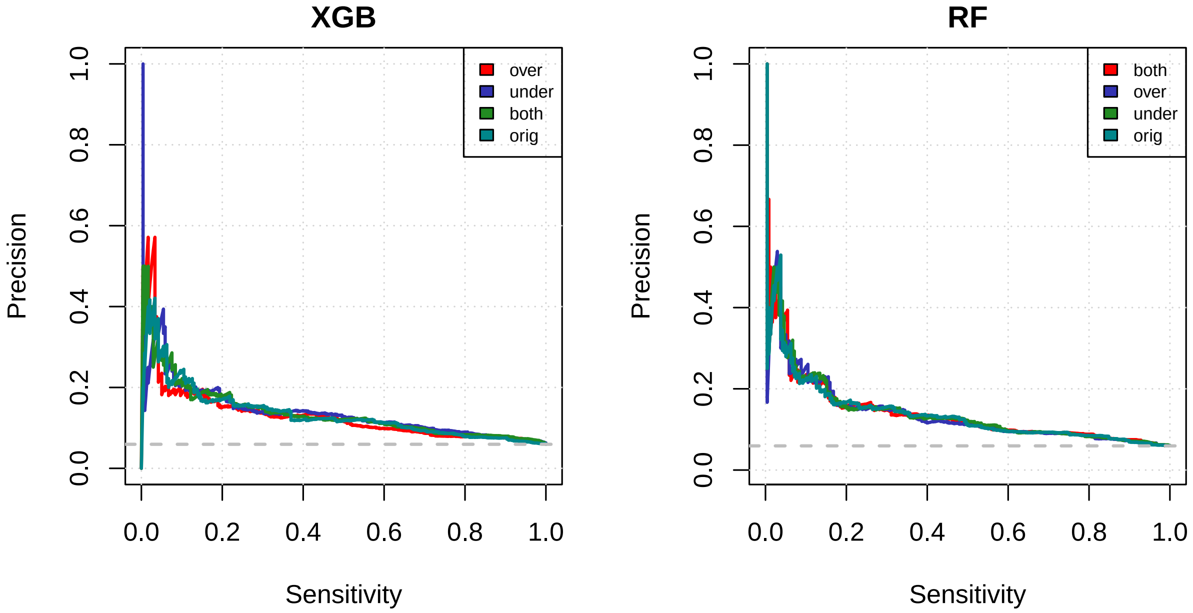

Figure 5: Forest and XGB resampled

Similar to the Glmnets the re-sampling does not yield increases in predictive performance as displayed in the plots above. Only for higher cut-offs, in the case of XGB, the trade-off between precision and sensitivity seems to be superior for the re-sampled models. However, for reasons previously outlined we are not interested in this range of cut-offs. Consequently, we may conclude that the re-sampling did not result in increases in model performance and thus the best performing model is still the elastic net utilizing the complementary-log-log link.

4.4 Best regressors

71 features are included in the optimal Elastic Net model with the cloglog link function. As the features have been discretionized in the data set, the variable importance is can be observed for each level of the feature, for example the PMOTSCO3 represents the 3rd level of the PMOTSCO feature. The 10 features with the highest explaining power are

- PMOTSCO: Contribution motorcycle/scooter policies
- PPLEZIER: Contribution boat policies
- ALEVEN: Number of life insurances
- AFIETS: Number of bicycle policies
- PPERSAUT: Contribution car policies
- PWAOREG: Contribution disability insurance policies
- APLEZIER: Number of boat policies
- MINKM30: Income $<$ 30.000
- PBRAND: Contribution fire policies

Overall, the important variables point towards customers that already own a number of other insurance policies and have a higher contribution to the policies. High levels of contribution to boat policies, car policies and disability insurance policies indicate a higher likelihood, that the customer would also be interested in a caravan insurance. Interestingly, a low level of boat insurance suggests an interest in caravan insurance as well, of course this may not be interpreted as a causal effect without further analysis.

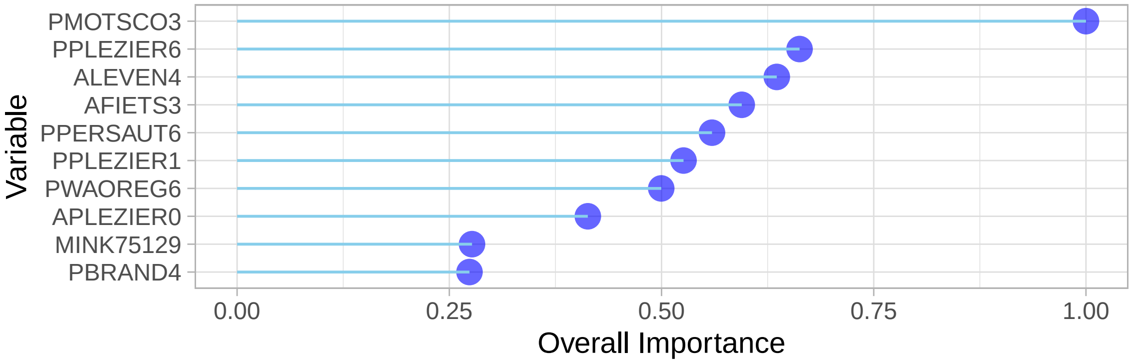

Figure 6: Variable importance in our best model

4.5 Comparison to the Kaggle Competition

In general, our results differ quite a bit from the winners of the Kaggle Competition as there was one winner for the description section and one winner for the prediction section. As the competition allowed to set a focus on a section, by selecting different winner, we decided to focus on the predictive task in our analysis. A general problem in the submitted solutions was that the participants often did not account for the unbalanced data.(Elkan 2001) This is closely linked to the other big problem identified with all the identified solutions, namely high variance. Again, this problem was often not recognized as such and often a more advanced approach in fine tuning was named as a potential improvement, which would have worsened the problem even further. (Van Der Putten and Van Someren 2004)

Charles Elkan, the winner of the prediction section used a Naïve Bayes learning approach and identified the best predictors with this strategy. The features with the largest explaining power in his model were a high purchase power class, private third-party insurance, boat policy, a social security insurance policy and a single fire policy with high contribution level. He would describe the most likely customer for a caravan insurance policy as a wealthier than average customer who already owns a car, which is insured with a high premium. (Elkan 2001)

YongSeog Kim and W. Nick Street, the winners for the description section, used a combination of artificial neural networks for predication with an evolutionary search. As they did not include all of the features provided, the predictive features were chosen by the evolutionary local search algorithm, which utilizes the quality and likelihood of neighboring individuals to experience the same feature. They found the feature on the number of and the contribution to car insurance to be the most important predictor. Additionally, an interesting insight was that highly wealthy customers are less interest in caravan insurance even though above average wealth is increasing the likelihood. (Kim and Street 2000)

The biggest similarity regarding the variable to our results is the importance of having a boat insurance in the predictors. Our regressors with the largest explaining power point into a similar direction as owning an above average number of insurance policies and having a high contribution to those is indicating a customer whose wealth is above average. Besides that, the features with the highest explaining power selected by our elastic net model were different ones.

The competition task was to hand in the 800 customers from the evaluation data set, who had the highest likelihood of being interested in obtaining a caravan insurance policy. There were 238 from the 4000 variables were caravan policy holders. The winning entry identified 121 correctly among his top 800 predictions. The next two best scores were 115 and 112. (Elkan 2001) Our approach led to correctly identifying 117 caravan policy holders among the top 800 predictions, assigning us the second place in the prediction task of the competition.

5 Economic operationalization

"[. . .]it is usually economically irrational to offer an insurance policy to some arbitrary percentage of customers. Instead, an offer should be made to a customer if and only if the expected profit from making the offer is greater than the cost of making the offer." (Elkan 2001)

This is how Charles Elkan, the winner of the Kaggel competition, highlights an important issues. In the context of predicting potential clients it is not about finding a fixed number of the customers who are most likely to accept an offer but to offer insurance to clients in a profit maximizing manner. To do so we will outline a simple model with constant profits per contract and costs per contact, α and β respectively. In this context α represents the profits of an insurance contract without the costs of contacting the client and β are just the costs connected with offering an insurance to a client. Then the expected per contact profit is given by,

$$\pi_i = \alpha p_i - \beta, \tag{1}$$

where p_i represents the probability of client i to contract an insurance. Offering an insurance is only profitable if (1) is non negative which transforms to

$$p_i \geq \frac{\beta}{\alpha}. \tag{2}$$

Equipped with this decision rule for offering insurances the total expected profit,

$$\Pi = \sum_{i=1}^{n} c.p.(\pi_i), \quad \text{where} \quad c.p.(\pi_i) = \begin{cases} \pi_i & \text{if} \quad p_i \geq \frac{\beta}{\alpha} \\ 0 & \text{else} \end{cases} \tag{3}$$

is maximal: making an offer to client i where $p_i < \frac{\beta}{\alpha}$ leads to a lower profit compared to (3); similarity not making an offer to client j where $p_j > \frac{\beta}{\alpha}$, i.e. choosing a decision rule such as $p_j > \tau$ with $\tau > \frac{\beta}{\alpha}$, leads to a lower profit compared to (3).

We thus see that choosing the cut off point equal to $\frac{\beta}{\alpha}$ maximizes profits in the frame of the stylized model. Given the parameters α and β this task is trivial, finding out about the actual values of these parameters however is quite challenging without company intern knowledge and is therefore out of the scope of this paper.

The approach above however gives a simple rule to determine the cut off point such that potential customers are selected in a profit maximizing manner. Given the estimates of the probability that a customer would accept an obtained offer and knowledge about their cost structure insurance companies could thereby easily target their customers optimally.

To illustrate the idea on the basis of our estimates assume that an insurance agent gets paid a fixed sum of $\alpha = 76.38$ for each signed contract. The costs connected with offering a contract to each customer are given by $\beta = 4.82$. Consequently the profit optimal cut off point is given by $cop = \frac{\beta}{\alpha} \approx 0.063$. Applied on the estimates of our best performing model this lead to the following confusion matrix.

Table 1: Confusion Matrix given the optimal cop (actual labels as columns)

	FALSE	TRUE
FALSE	2657	87
TRUE	1105	151

As we see above, offering in the described way leads to a high number of false positives as the costs of contacting a client that does not accept the offer are quite low compared to the opportunity costs of not offering a client that would have accepted an offer.

6 Conclusion

Predicting demand for Caravan insurances in the frame of the CoiL Challenge led to a variety of interesting results.

Given the insight that (by the economic structure of the problem) we are interested in predictions based on low cut off points we find that comparably simple methods like Glmnets outperform more complex models such as random forest or XGB models. Under the models of consideration a Glm elastic net with a complementary log-log link function led to the best results. Furthermore, the driving explanatory factors are, beside income, other policies. That is information about consumer decisions for similar products is more important than socio-demographic features. An insight that would be worth investigating in a more general setting. Finally, we outlined that accounting for the imbalancedness of the data did not lead to an improvement of the considered models.

Based on these results (Glmnet with unengineered data led to the best predictions) we conclude that a relatively simple solution is the best a approach to the outlined problem. However, it has to be mentioned that we had a 20 year technological advantage on the competion's participants. For instance, elastic net penalization was only introduced by Zou and Hastie (2005) a few years after the competition. Similarly, random forests were only introduced in 2001. So in some sense, we simply had a larger toolbox at our disposal.

References

Elkan, Charles. 2001. "Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000." *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July. https://doi.org/10.1145/502512.502576.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://www.jstatsoft.org/v33/i01/.

Gupta, Rohan. 2019. "An Introduction to Discretization Techniques for Data Scientists." December 6, 2019. https://towardsdatascience.com/an-introduction-to-discretization-in-data-science-55ef8c9775a2.

Kim, Y, and WN Street. 2000. "CoIL Challenge 2000: Choosing and Explaining Likely Caravan Insurance Customers." *Technical Report 2000-09.*

Van Der Putten, Peter, and Maarten Van Someren. 2004. "A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000." *Machine Learning* 57 (1): 177–95.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20. https://doi.org/https://doi.org/10.1111/j.1467-9868.2005.00503.x.