Caravan Insurance: Selling Efficiently

Where should marketing and sales be focused?

Erik Carrion

2023-11-20

Executive Summary

The caravan insurance challenge asked competition entrants to use a training set of 5800 observations in 85-D space to develop a predictive model that provides insight into the factors that positively influence the sale of caravan insurance.

The present analysis considers how 3 linear models and a tree based method perform in an expanded feature space. After one-hot-encoding the original feature space, we go from an 85-D to a 538-D sparse feature space. In this higher dimensional space, Elastic-Net performs the best, yielding a parsimonious model of 58 predictors.

We find that if the company wishes to increase their sales of caravan insurance they should focus their marketing & sales efforts on:

```
1. Areas with higher proportions of:
    - middle class families,
    - car ownership,
    - religious diversity, and
    - religious activity.
2. Where the company has active:
    - contribution car policies,
    - contribution fire policies, and
    - boat policies
```

They should place special attention to areas where they have active boat policies. If a neighborhood can support the expense of boat ownership, then the expense of caravan ownership, which includes caravan insurance, is within reach.

Introduction

The purpose of this analysis is to compare how elastic-net, ridge, lasso, and random forest compare in a classification setting where the dimensionality of the predictor space is very high and we're working with a moderate amount of data.

To this end we make use of caravan insurance data that was provided as part of the CoIL 2000 Challenge on Kaggle. We start with a description of the data before proceeding with an exploratory analysis to gain a better understanding of the company's customers.

To assess model performance, we employ a 50 run simulation where we train and validate all 4 models using a 90/10 split. At each iteration we record train and test AUC and the time it takes to fit each model.

Once the simulations have completed, we cross validate each model on the entirety of the training data and inspect the resulting model to gain insight into the factors that drive caravan insurance sales.

Data Description

The dataset used contains information on the customers of an insurance company. It includes demographic data and product usage data at a zip-code level of resolution. Customers can be classified according to 1 of 10 main types each of which can be described by 1 to 5 different customer sub-types allowing for 40 unique customer combinations.

For example, Successful Hedonists are characterized as older, affluent individuals with status while customers categorized as Living Well are characterized as younger individuals enjoying apartment living in a culturally and economically diverse urban area.

Customer data is either ordinal or nominal while demographic & product usage data is all ordinal. Demographic data is recorded as the *percentage* of the given variable observed within the given zip code while product usage data is recorded as the *total* of the given variable observed in the given zip code.

Customer Characteristics

Main Types

Customers can be assigned one of ten main customer types:

```
1. Successful hedonists
2. Driven Growers
3. Average Family
4. Career Loners
5. Living well
6. Cruising Seniors
7. Retired and Religious
8. Family with grown ups
9. Conservative families
10. Farmers
```

The data dictionary only provides the labels described above. Without the precise definitions of the main customer types, our inferences will be limited.

We see they are descriptive, specific, and have a touch of linguistic flourish - 'cruising seniors' & 'living well', for example. This stands apart from the mathematical precision we expect from an insurance company's core business, suggesting the labels were developed by a research/marketing team either internally or externally.

The specificity of the labels & the fact the dataset was provided through Kaggle, a website focused on machine learning competitions, suggests they were developed using a data-driven approach. If so, we can infer the labels describe the entire population of the company's customers.

The combination of main-type and sub-type allows us to see how the sub-types are ascribed to the main-types. For each main type there are 2 to 5 associated sub-types and except for sub-type 5,‘mixed seniors’, every sub-type is associated with a single main-type.

	1	2	3	4	5	6	7	8	9	10
1	124	0	0	0	0	0	0	0	0	0
2	82	0	0	0	0	0	0	0	0	0
3	249	0	0	0	0	0	0	0	0	0
4	52	0	0	0	0	0	0	0	0	0
5	45	0	0	0	0	0	141	0	0	0
6	0	119	0	0	0	0	0	0	0	0
7	0	44	0	0	0	0	0	0	0	0
8	0	339	0	0	0	0	0	0	0	0
9	0	0	278	0	0	0	0	0	0	0
10	0	0	165	0	0	0	0	0	0	0
11	0	0	153	0	0	0	0	0	0	0
12	0	0	111	0	0	0	0	0	0	0
13	0	0	179	0	0	0	0	0	0	0
15	0	0	0	5	0	0	0	0	0	0
16	0	0	0	16	0	0	0	0	0	0
17	0	0	0	9	0	0	0	0	0	0
18	0	0	0	19	0	0	0	0	0	0
19	0	0	0	3	0	0	0	0	0	0
20	0	0	0	0	25	0	0	0	0	0
21	0	0	0	0	15	0	0	0	0	0
22	0	0	0	0	98	0	0	0	0	0
23	0	0	0	0	251	0	0	0	0	0
24	0	0	0	0	180	0	0	0	0	0
25	0	0	0	0	0	82	0	0	0	0
26	0	0	0	0	0	48	0	0	0	0
27	0	0	0	0	0	50	0	0	0	0
28	0	0	0	0	0	25	0	0	0	0
29	0	0	0	0	0	0	86	0	0	0
30	0	0	0	0	0	0	118	0	0	0
31	0	0	0	0	0	0	205	0	0	0
33	0	0	0	0	0	0	0	810	0	0
34	0	0	0	0	0	0	0	182	0	0
35	0	0	0	0	0	0	0	214	0	0
36	0	0	0	0	0	0	0	225	0	0
37	0	0	0	0	0	0	0	132	0	0
38	0	0	0	0	0	0	0	0	339	0
39	0	0	0	0	0	0	0	0	328	0
40	0	0	0	0	0	0	0	0	0	71
41	0	0	0	0	0	0	0	0	0	205

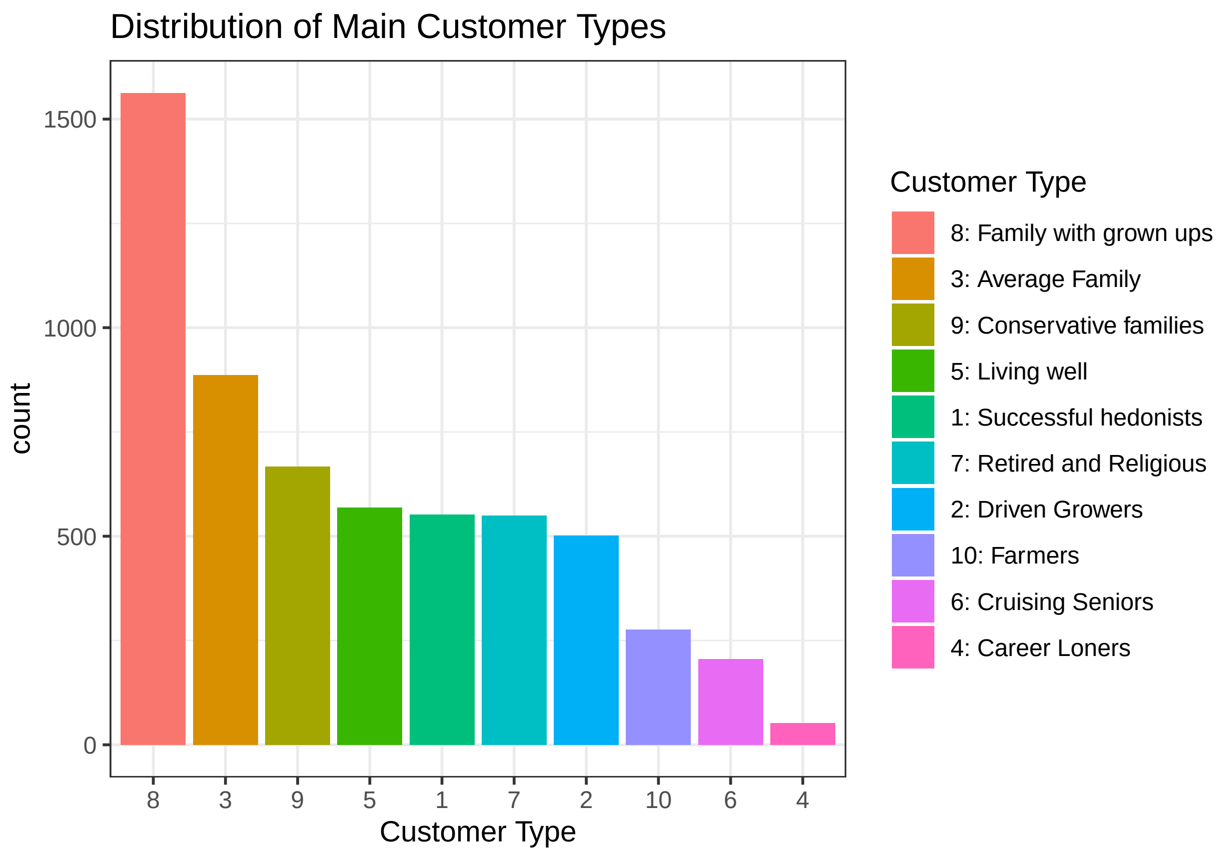
We see that ‘Career Loners” (main type 4) are the least prevalent of the groups while ’Middle Class Families’ (main type 8) are the most prevalent.

```
##    4    6   10    2    7    1    5    9    3    8
##   52  205  276  502  550  552  569  667  886 1563
```

The associated sub-types for each main-type allows us to better understand how the main-types are defined. Of the ten main types, six are family related, two are related to conservative values, and four are related to seniors.

```
## $`Successful Hedonists`
## [1] "High Income, expensive child" "Very Important Provincials"
## [3] "High status seniors"          "Affluent senior apartments"
## [5] "Mixed seniors"
##
## $`Driven Growers`
## [1] "Career and childcare"  "Dinkis"                "Middle class families"
##
## $`Average Family`
## [1] "Modern, complete families" "Stable family"
## [3] "Family starters"           "Affluent young families"
## [5] "Young all american family"
##
## $`Career Loners`
## [1] "Senior cosmopolitans"      "Students in apartments"
## [3] "Fresh masters in the city" "Single youth"
## [5] "Suburban youth"
##
## $`Living Well`
## [1] "Ethnically diverse"       "Young urban have-nots"
## [3] "Mixed apartment dwellers" "Young and rising"
## [5] "Young, low educated"
##
## $`Cruising Seniors`
## [1] "Young seniors in the city" "Own home elderly"
## [3] "Seniors in apartments"     "Residential elderly"
##
## $`Retired & Religious`
## [1] "Mixed seniors"                    "Porchless seniors: no front yard"
## [3] "Religious elderly singles"        "Low income catholics"
##
## $`Family with Grown Ups`
## [1] "Lower class large families"   "Large family, employed child"
## [3] "Village families"             "Married with children"
## [5] "Mixed small town dwellers"
##
## $`Conservative Families`
## [1] "Traditional families"    "Large religous families"
##
## $Farmers
## [1] "Large family farms" "Mixed rurals"
```

Looking at the distribution of main-types, the top 3 main types make up 53% of observations and are all family related.

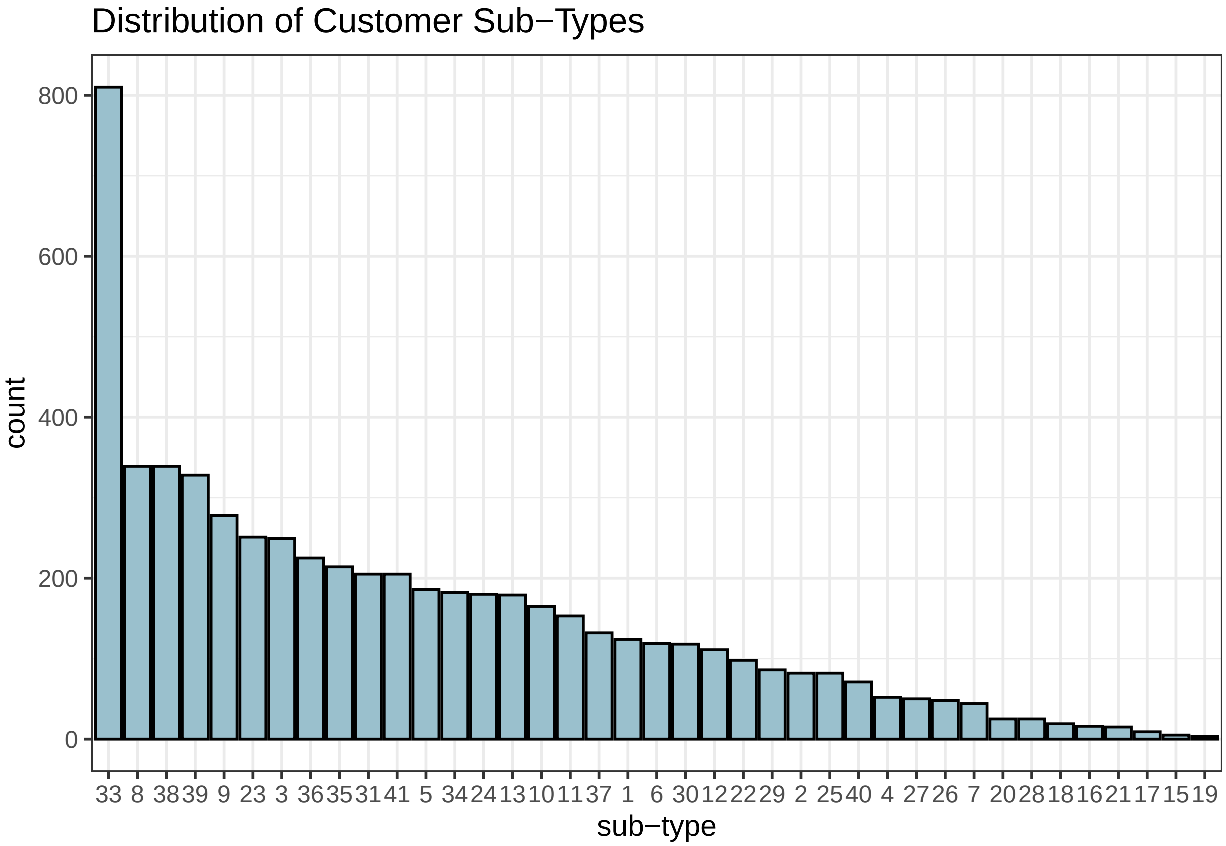


Sub-Types

The data dictionary list 41 different customer sub-types. After accounting for duplicates and omitted entries, we are left with 39 different customer sub-types. We can view these sub-types as granular descriptions of 3 overarching sub-types defined by their lowest common denominator: seniors, families, and individuals. Of the 40 sub-types, seniors represent 25% of types, while families and individuals each each make up 37.5% of labels.

```
##    sub_type                      label count       prop   cumprop
## 1        33 Lower class large families   810 0.13912745 0.1391274
## 2         8      Middle class families   339 0.05822741 0.1973549
## 3        38       Traditional families   339 0.05822741 0.2555823
## 4        39    Large religous families   328 0.05633803 0.3119203
## 5         9   Modern, complete families   278 0.04774991 0.3596702
## 6        23            Young and rising   251 0.04311233 0.4027825
## 7         3        High status seniors   249 0.04276881 0.4455514
## 8        36      Married with children   225 0.03864651 0.4841979
## 9        35           Village families   214 0.03675713 0.5209550
## 10       31        Low income catholics   205 0.03521127 0.5561663
```

Continuing the trend we saw with the main customer types, the sub-types are dominated by family related labels. Of the top 10, 7 are family based categories. In combination with the most prevalent main customer types, we can infer that, at least based on the dataset, that their customers are primarily defined by their family affiliation/structure. Their largest customer sub-type is large lower class families which are part of main-type 8, ‘Family with grown ups’, which also includes Lower class large families, Large family, employed child, Village families, Married with children, Mixed small town dwellers so we can infer they are unlikely to be affluent. If they aren't defined in relation to a family, they are either on their way to the top ('Young and rising') or already there ('High status seniors').



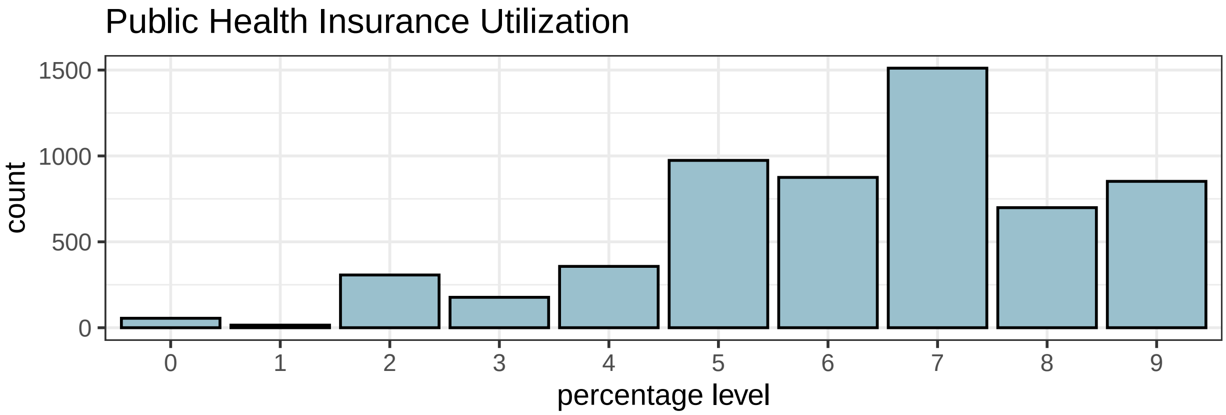
Predictors

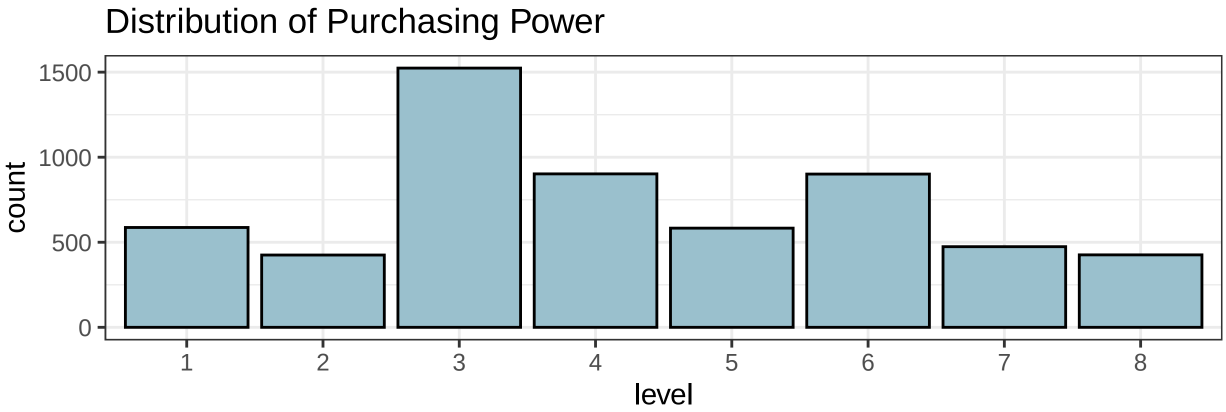
We have 85 predictors all of which are either nominal or ordinal. The ordinal variables represent percentages or totals and each can take on one of ten values from 0 to 9. Some variables do not make sense in this context. For example, for Average Income, MINKGEM, the most common values are 3 and 4 which correspond to 24-36% and 37-49% respectively. The data dictionary only says these numbers relate to the percentage observed in the given zip code. Without a reference baseline, the statement 'average income is 24-36%' doesn't confer a usable meaning.

Looking at the ordinal predictors only, we can use their overall sums as a starting point for an investigation into their distributions. Variables with larger sums can be expected to have many observations assigned to higher levels and vice-versa. Below are the top 10 variables by sum.

```
##         top top10   bottom bot10
## 1   MZFONDS 36545  AZEILPL     3
## 2    MRELGE 36000  PZEILPL     5
## 3     MAUT1 35167  AVRAAUT    13
## 4    MHKOOP 27781  AWAOREG    27
## 5    MGODPR 26938 APERSONG    31
## 6  MOPLLAAG 26621 APLEZIER    35
## 7  MFWEKIND 25036   AWERKT    36
## 8    MHHUUR 24667  AGEZONG    38
## 9  MKOOPKLA 24664  AINBOED    46
## 10  MINKGEM 22033  PVRAAUT    55
```

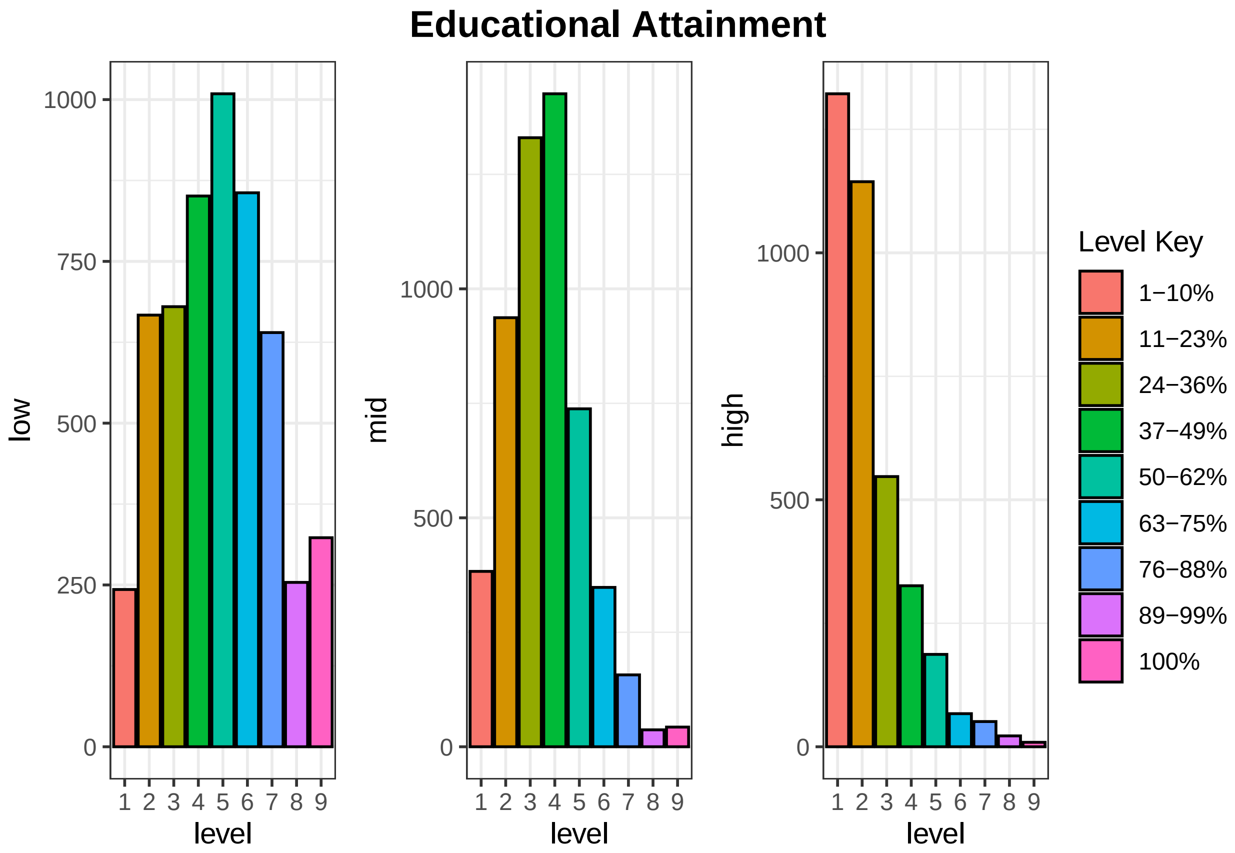
We'll compare the distributions of MZFONDS to MKOOPKLA to see how they relate. MZFONDS represents the percentage of the zip code utilizing public health insurance and MKOOPKLA refers to the purchasing power class the customer belongs to. The former is measured as the "percentage of each group, per postal code" and the latter is measured on a scale of 1 to 8. We see that the majority of customers are assigned to a level of 5 or above for public health utilization and most customers are below a 5 in purchasing power class. Together, it supports our initial inference that most customers are not likely to be affluent.





Education is encoded using 3 multinomial indicator variables for low, medium, and high educational attainment. Inspecting the distribution of educational attainment, we observe that attainment is not distributed comparably between the three levels and within the levels we observe differences in the patterns of variation.

```
##   level percentage  low  mid high total     low.p     mid.p     high.p
## 1     1       1-10%  243  383 1322  1948 0.1247433 0.1966119 0.67864476
## 2     2      11-23%  667  937 1144  2748 0.2427220 0.3409753 0.41630277
## 3     3      24-36%  680 1330  547  2557 0.2659366 0.5201408 0.21392257
## 4     4      37-49%  851 1426  326  2603 0.3269305 0.5478294 0.12524011
## 5     5      50-62% 1009  738  187  1934 0.5217166 0.3815926 0.09669080
## 6     6      63-75%  856  348   67  1271 0.6734854 0.2738002 0.05271440
## 7     7      76-88%  640  157   51   848 0.7547170 0.1851415 0.06014151
## 8     8      89-99%  254   37   22   313 0.8115016 0.1182109 0.07028754
## 9     9        100%  323   43    9   375 0.8613333 0.1146667 0.02400000
```



The distribution of customers with low educational attainment shows a somewhat symmetric distribution about level 5. For customers with a mid level of attainment, the distribution is more right skewed than that of low educational attainment, peaking at level 4. High educational attainment is extremely right skewed, peaking at level 1. Overall mid and high educational attainment is predominantly observed in lower percentages while low educational is predominant in the higher percentage levels.

Given that the company's customers tend to have lower educational attainment and occupy lower purchasing power classes, we now look to the distribution of income. The data dictionary defines 4 indicator variables for income with as stratified and indicated by the variables MINKM30, MINK3045, MINK4575, MINK7512, and MINK123M which separates income into 5 levels with breaks at $30k, $45k, $75k, and $122k per annum. Given what we've gathered so far, what we see in the income distribution is not unexpected: the company is less likely to have customers located in high income zip codes.

```
##    level INKM30 INK3045 INK4575 INK7512 INK123M
## 1      0   1304     465     891    3246    4900
## 2      1    630     268     657    1359     763
## 3      2   1094     919    1165     736      96
## 4      3   1079    1147    1215     246      36
## 5      4    599    1356    1034     147      24
## 6      5    568     931     498      71       1
## 7      6    293     406     125       8       0
## 8      7    156     205      93       1       1
## 9      8     48      35      53       4       0
## 10     9     51      90      91       4       1
```

Model Discussion

The present analysis compares 4 different models: elastic-net, ridge, lasso, and random forest using AUC as our performance metric of choice.

Ridge, lasso, and elastic-net are all linear penalized/regularized regression models. Lasso and Elastic-Net provide a level of variable selection while ridge maintains all variables in the final model. If there are non-linearities in the original feature space, a linear model will not serve well. To ameliorate this issue, we can work work in higher dimensions.

Since this is a classification problem, we use the binomial for the model family in glmnet. This choice drives our decision to one hot encode the feature matrix. This increases the number of features from 85 to 538. Linear models developed in 538-D space will approximate non-linear relationships, to the extent they exist, in 85-D space reasonably well.

In addition to the linear models, we also employ the random forest algorithm to build a prediction tree in this higher dimensional space. Despite its computational expense, the random forest algorithm is widely used and known for yielding accurate models.

We compare the 4 models using AUC and their respective run times to identify which of the four models is the best among the 4. We do not consider other algorithms which may be better suited to the problem at hand like logistic regression or support vector machines. Our purpose here is to investigate the effectiveness of using linear methods in very high dimensions to model relationships in lower dimensional space.

Model Validation

To assess model effectiveness, we fit the four models on ninety percent of the data, retaining ten percent for validation. We employ cross-validation on the linear models to determine the optimal amount of regularization to apply to the penalty parameters. Random forest, due to its construction, doesn't require cross validation. We repeat this 50 times and assess performance using aggregate measures.

After we've performed the 50 simulations, we fit a final model on the entire training set before proceeding to assess their performance on the holdout set.

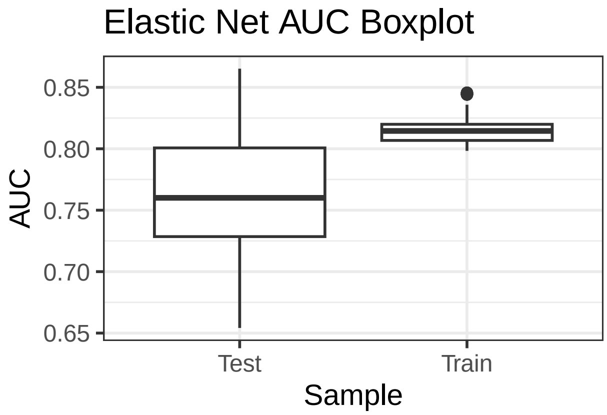
Model Results

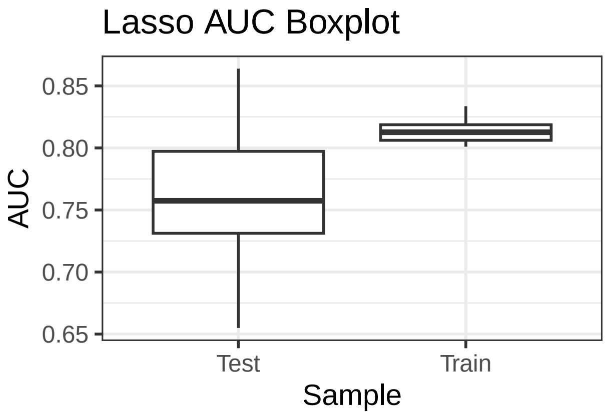
Each of the 50 simulations took an average of 3 minutes per iteration. When we compare the three linear models to the tree-based model, we see that in terms of training time, the three methods which perform a sort of variable selection each took over 1 minute to train each compared to 40 seconds for Ridge which uses all the variables in the final model.

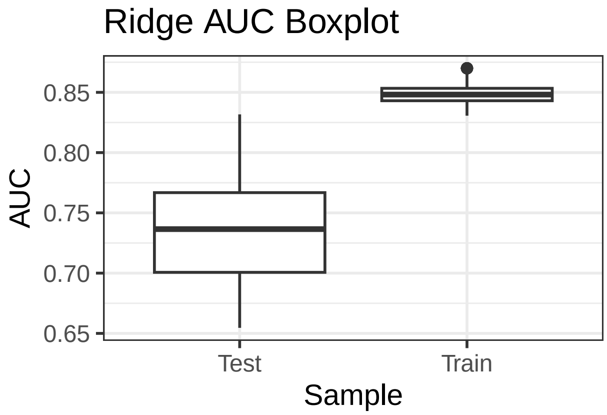
In terms of performance, elastic-net & lasso performed the best, followed by ridge and random forest. Elastic Net just barely beats out lasso while taking a few seconds less to fit.

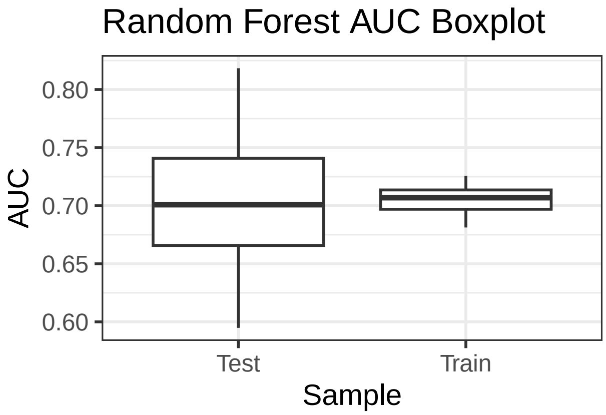
```
##          Method       AUC  Time
## 1   Elastic Net 0.7600801 62.34
## 2         Lasso 0.7573914 65.97
## 3         Ridge 0.7365212 40.27
## 4 Random Forest 0.7009719 62.42
```

When we plot the distribution of AUC over the 50 iterations, we get the box plots below. We see that for the training data, the IQR is significantly tighter than for the test data. In each case, except for random forest, training performance is significantly better on average than on the test data. Finally, all three linear models perform significantly better than the tree based method.


Elastic Net AUC Boxplot


Lasso AUC Boxplot

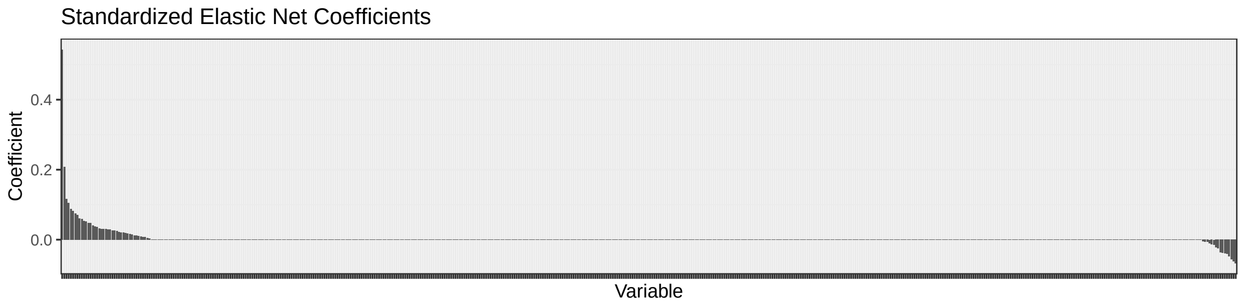

Ridge AUC Boxplot

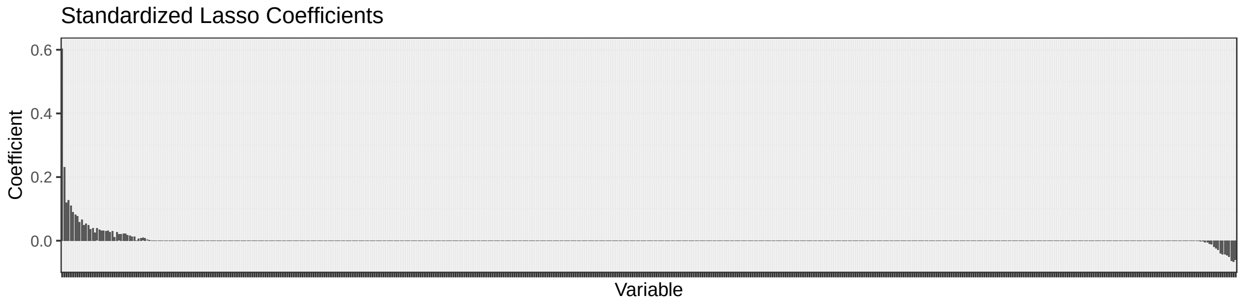

Random Forest AUC Boxplot

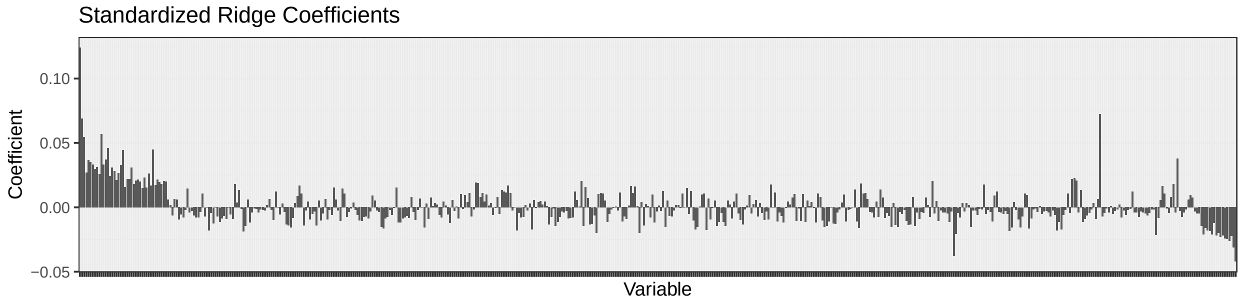
Elastic-Net has the highest median AUC out of all four models.

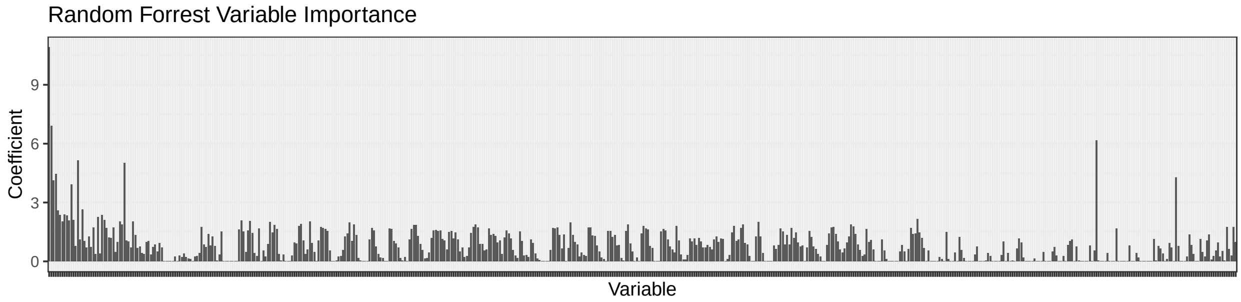
Coefficients

Do the methods that perform variable selection agree with each other? While we expect some variation, if the data sufficiently captures the relationship between the response and the predictors, then we expect to see general agreement among the four methods. To do so visually, we plot the standardized coefficients and then order them according to their importance as determined by random forest.









Lasso and elastic net perform comparably with respect to variable selection. They roughly choose the same variables while generating very similar coefficient estimates. For ridge and random forest we observe a relation between the ridge coefficient estimates and the variable importance reported by random forest.

Overall, all four methods agree in terms of which variables are most important and each achieves a median AUC of at least 70%, with elastic-net performing the best out of the four.

Top-5 Variables

All 4 models agree that ‘PPERSAUT6’ & ‘PBRAND4’ are significant variables. They correspond to levels in PPERSAUT & PBRAND which relate the number of:

```
1. contribution car policies:
    - all agree that level 6 is the most important
```

```
2. contribution fire policies
    - all 4 agree that level 4 is significant
    - 2 out of 4 models also include level 3
```

Specifically, each model agrees that postal codes with 1000-4999 contribution car policies & 200-499 contribution fire policies are significant for determining who will and who will not purchase caravan insurance.

Overall, car insurance (PPERSAUT) and fire insurance (APERSAUT) are chosen by every model. They are in the top 5 for ridge and random forest and in the top 15 for lasso and elastic net.

Despite being observed in only 31 of the 5,820 observations, all three linear models agree that APLEZIER1, a positive number of boat policies, is a significant variable. Given its relatively low prevalence in the data, special focus should be placed on areas where the company has a positive number of active boat policies.

We note that multiple levels of PBRAND are found to be significant across the four models. Further, we see that the customer main type, 'middle class families', is found in 2 of the 4 coefficient sets. Finally, random forest places higher significance on third party insurance products.

```
##   number     elnet     lasso     ridge random_forest
## 1      1 PPERSAUT6 PPERSAUT6 PPERSAUT6     PPERSAUT6
## 2      2   PBRAND4   PBRAND4 APERSAUT1       PBRAND4
## 3      3 APLEZIER1   PBRAND3   PBRAND4     APERSAUT1
## 4      4   PBRAND3 APLEZIER1 APERSAUT2      PWAPART2
## 5      5  MOSTYPE8  MOSTYPE8 APLEZIER1      AWAPART1
```

Bottom-5 Variables

At the bottom end we see that MINKGEM, PBRAND, MAUTO, and MGODOV are all at the bottom 5 for elastic net, ridge, and lasso. For random forest we first require a positive mean decrease in gini before inspecting the bottom 5 variables, resulting in some disagreement with the linear models.

```
##   number      elnet      lasso     ridge random_forest
## 1      1    PBRAND2   MINKGEM3   PBRAND2      PWABEDR1
## 2      2   MINKGEM3 MOSHOOFD10 PPERSAUT5       PWERKT6
## 3      3 MOSHOOFD10    PBRAND2  MINKGEM3        MAUT08
## 4      4     MAUT04     MAUT04    MAUT04       PWERKT2
## 5      5    MGODOV1    MGODOV1   MGODOV1       MGODGE8
```

Overall, if we observe the following conditions we can expect a lower likelihood of caravan insurance being purchased:

```
- moderately low (level 4) or high (level 8) levels of *no* car ownership
- low numbers (level 2) of fire policies
- low levels of 'other' religious affiliations
- high levels of 'no' religious affiliations
- whether a customer is a farmer
- whether average income falls in to the 24-36% bucket
```

Conclusion

The models suggest that focusing marketing and sales efforts in postal codes that meet the following criteria will maximize the likelihood of selling caravan insurance:

```
1. Has 1000-4999 active contribution car policies,
2. Has 200-499 active contribution fire policies,
3. Has 1-49 active boat policies
4. Higher proportions of middle class families
5. Higher rates of religious activity
6. Higher rates of religious diversity
6. Higher rates of car ownership
```

Of the four models tested, elastic-net performed the best, using a mix of the L2 and L1 penalty and, like lasso, yields a parsimonious model with only 58 predictors. Run time was not a significant consideration in this scenario. Given its overall performance and parsimony, we recommend elastic-net over the other three models.

Special focus should be placed on areas where the company has greater than 0 active boat policies. Boat ownership implies an area with excess disposable income. If they can afford the expense of a boat and a family, a caravan is within reach and thus, so is caravan insurance.

References

1. P. van der Putten and M. van Someren (eds) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000.
2. Data: https://github.com/erikscarrion/caravan-insurance/blob/edx_ds_cert/caravan-insurance-challenge.csv