Coil Data Predictive Model Project

Matthew DeGiacomo

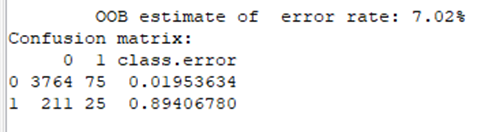
SNHU DAT-640 Predictive Analytics

The company we are focusing on in this data mining project is The Insurance Company (TIC) Benchmark. This company is an insurance company who has a large suite of different types of insurances stemming from life insurance, to boat, property, car, and the one we are trying to promote, caravan insurance. The company is looking to expand their business, specifically the caravan insurance policy, but the company has not been engaging in efficient marketing tactics to build and expand on their customer base. The company’s problem currently is that they are wasting marketing dollars on sending out letters to all their customers blindly with no real strategy behind it. Customers get the same letter, which has gone out to everyone, they think its junk and throw them away, the company wastes its marketing budget, and landfill sites get filled with all the unread mail. The companies problem can be described as they do not have an effective way of reaching their customer base to introduce new products or have enough information on their customer’s to determine which customers should be offered the specific products that fit their needs. The research questions can be states as, **“What model is the most appropriate for maximizing the customer information dataset to provide actionable insights on customers who would be more inclined to buy caravan insurance?”** The main limitation is the company does not have any in depth knowledge of their customer base because they have no way of translating their current customer information to determine which prospects to target. Through this analysis, the company hopes to have a solid knowledge of which customers to target for their caravan insurance based on their purchase history and demographics. The potential value that creating a predictive analytical model will bring to this organization is limitless. If a predictive model was created and implemented for determining the qualities of certain customers who would be more likely to buy caravan insurance, the company would save money on marketing dollars, reduce waste while also promoting themselves to be greener, and then potentially selling more caravan insurance policies which is what they originally set out to do. The long term potential value is immeasurable as well because you can then take that model and use it towards selling all different and new products, next it can be boat insurance or motorcycle insurance and so on. You can take those saved marketing dollars and narrow your targeted customer base to ones that are more likely to buy any of the products of insurances offered. You can take what the company was spending to promote caravan insurance before and now that could be split into a couple different products meaning what you used to spend to market for one product you can spend that same amount to market multiple products with also a higher chance of selling them because you have narrowed down a target group who would be more likely to get that specific insurance. In the overall business, this can increase their sales, increase marketing efficiency, set up the company to be more efficient in their marketing spending, reduce landfill and paper output, potentially expand their customer base, and give their sales people stronger leads to close on which creates better office moral.. It is to be noted though that these predictive models won’t give you the answer to your problems but more pointing you in the right direction and giving you outputs that can possibly lead to your answer you are looking for.

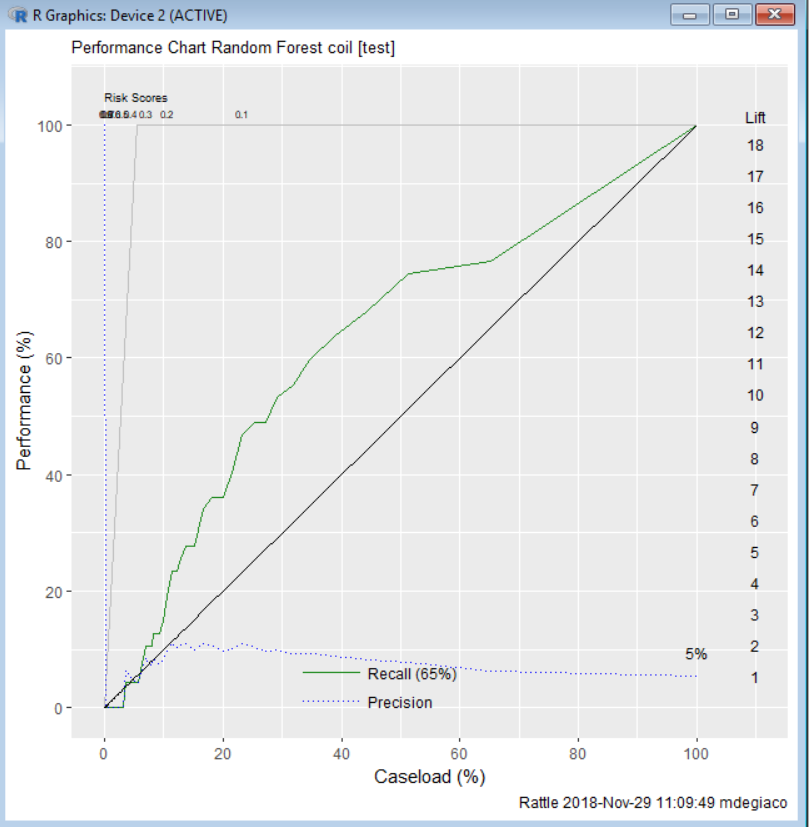
The specifications I need to follow to select a predictive algorithm are: what kind of data do I have: a mix of numeric and categorical data, what am I trying find out: customer retention, recommendations to products, customer insights, and lastly marketing effectiveness. When trying to find out customer retention/customer insights, the classification algorithms work best, like a random forest or decision tree. You can use classification algorithms that create rules to decide the outcome of a given scenario, which can help answer questions like: “Is this customer likely to respond to our marketing campaign?”, Given the specifications laid out and what the Insurance companies goal is in regards to developing a predictive model, I would recommend a classification model, more specifically a random forest model. Using my coil data set with variable #86 CARAVAN Number of mobile home policies, I ran a standard tree and then a random forest with this data and my results were interesting in the sense that going forward I’m going to run with the random forest for my project. When I did the standard tree my results were very simple, not enough results rather, only 1 rule, Rattle told me it only consists of one node so a tree couldn’t be made, and it was accurate though at a 95% accuracy. The reason I chose a random forest is due to the amount of variables that go into each answer of the reasons why a person is more likely to have or not have caravan insurance, it would take multiple tree’s to run different variables with each other. There is not a simple direct answer so the standard tree is already at disadvantage for getting the best answer and that’s the random forest is better for this type of data that I have because so different variables can be used to get the same answer and sometimes in getting the same answer completely different ones were used that don’t overlap. So running multiple trees with multiple different styles of variables is the best way to maximize the data for this set.

The proper analytic tools to help us facilitate the implementation of my random forest model is R and Rattle. R is what this class was essentially all about, it’s a free open source software environment for stats, computing, visuals, modeling, and many other functions. It can run of UNIX, Windows, MAC Os and is categorized as a predictive analytic software. Within R there is a package that you can download that enhances your modeling experience even more called Rattle. Rattle allows the user to quickly load data from multiple sources, transform and explore the data, build and evaluate models, and export models as PMML or as scores. Little to no code is needed by the user to interact with Rattles interface, but what’s great about Rattle is it keeps a log of the code from the actions you performed and you can export that code to use for a later date to use in R. Rattle is where all my work for this project was done, it’s easy to use, provides excellent insight with powerful tools, and can be done with just clicks of a button. R and Rattle can do everything from cleansing your data, exploring the data, build a model, evaluate that model, visualize the model, and improve the model. R and Rattle will help give you scores for the customer, in this project, the scores represent their likeliness to buy caravan insurance. Besides classification models, R can also do linear and nonlinear modeling, stats tests, time series analysis, clustering, interactive visualizations, and the limits for R start and end with the user. Besides all the functions that R and Rattle have to actually do the work needed to create and implement the model, they are free, so it cost’s the company no money to download, and it is easy to use and understand. Along with the data scientist who is doing the modeling, you can teach some of the other higher up’s in the company the basics in Rattle to be able to pull data themselves or at the least have them understand the outcomes you came up with.

The method that I used to score my model is the one we learned about in the book using Rattle. After my random forest model is created, I went to the Evaluate Tab 🡪For type I clicked on Score 🡪 Model is the Random Forest 🡪 For data I am testing the model vs the Test Data 🡪 For the report I did one for class and included all variables and then I did another CSV score with probability and included all variables. The outputs the scores gave me were either 1 or 0, indicating if the customer bought caravan insurance (1) or not (0). The score with the test data gave me 25 for the output of 1. What is interesting about this is my confusion matrix for my model has the number 25 as the amount of true positives (predicting they will buy and they did) and the scores provided have 25 outputs as 1. I selected this method because the testing data is the data we use to test our final model and I can tell its accurate predicting that 25 people will buy and that was the same as my true positive count for my model. Looking at my final random forest model, I made some adjustments in the parameters during milestone 5 that improved my model: based off the error matrix I cut down the amount of trees from 500 to 100 and I added more variables into the forest after some trial and error I came to add a total of 50 to give me my best results. My OOB error rate is 7%, so my model is 93% accurate which is really good, 99% correct in predicting who is not going to buy caravan insurance, only 11% accurate in predicting who is going to buy caravan insurance but what this model does help with is narrowing down customers with a higher chance of buying, so being able to narrow down that 236/4075 customers that have a higher chance of buying shows value right there. With such a small amount predicted to buy, there’s a small margin for error in correctly predicting, so that 11% accuracy doesn’t look good but it doesn’t affect the overall models error rate that much.

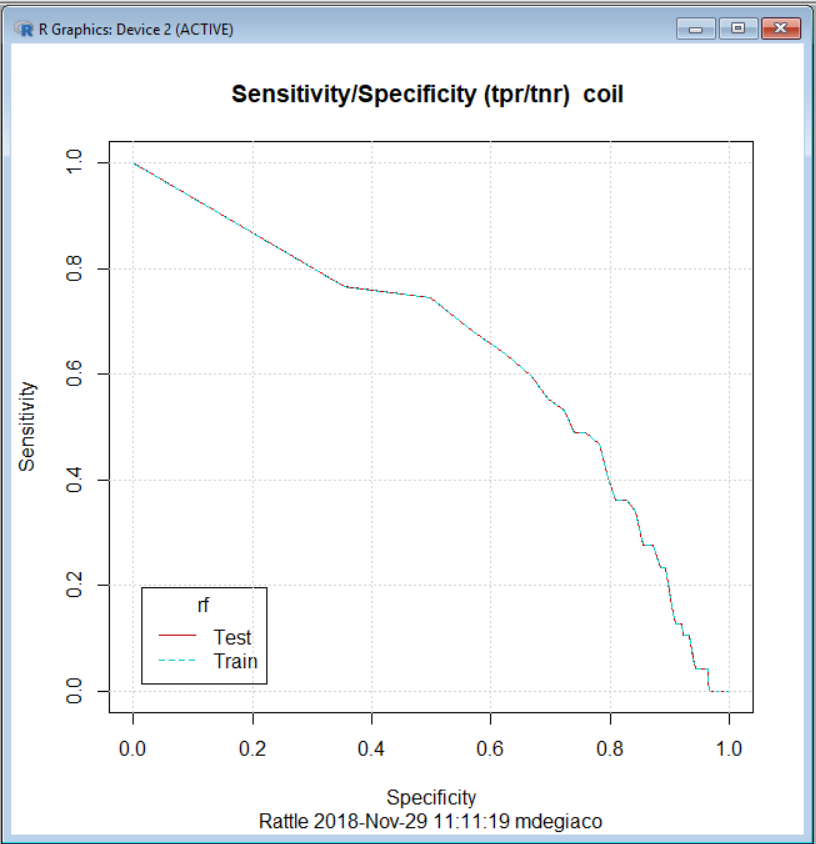


Looking at my risk chart I made with the test dataset, I have a strike rate of 5%, meaning that 5% of people are predicted to buy caravan insurance which lines up with my amount predicted to buy (236/4075 = 5%). What’s interesting about my strike rate line is that after 10% caseload, it stays very consistent only wavering 5% total, at 30% caseload the strike rate starts to decrease from its max strike rate of that 10% and it slows creeps down to 5% at 100% caseload. What this means is that even out of all the customers the insurance company has, there are just not a lot of people who are in the market to buy caravan insurance. Now comparing my random selection at 50% to my green score line at 50%, the random selection has 50% of the 5% strike rate at 50% caseload but the green score line is at 75% of the 5% strike rate at 50% caseload, and not hitting 90% until about 85% caseload.

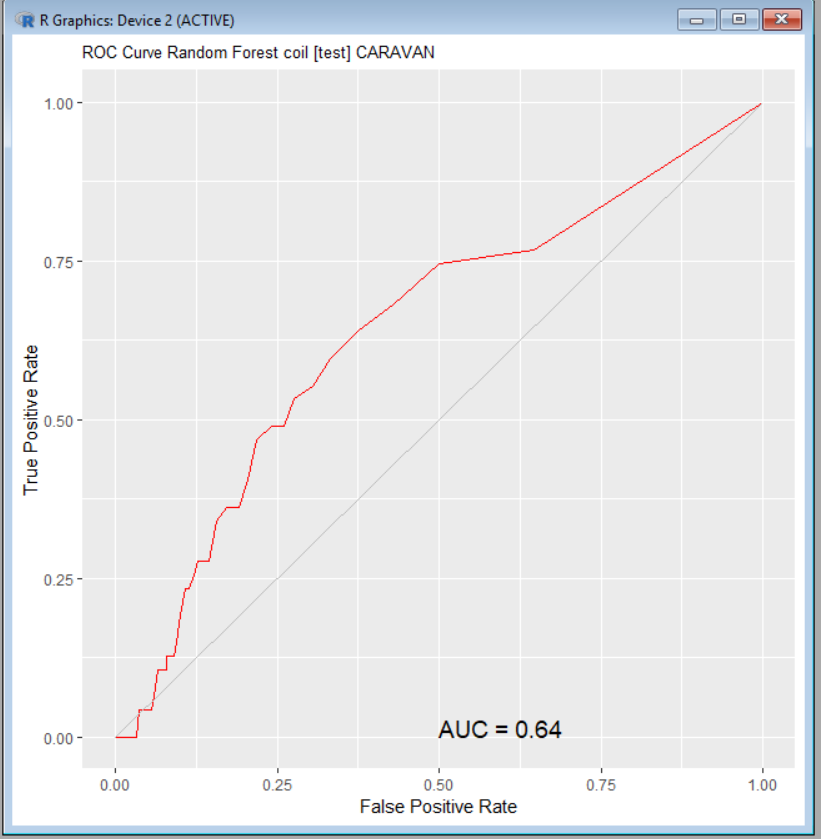


Now taking a look at a couple more charts, staring with a Sensitivity (True positive rate) vs Specificity (True Negative Rate) with the test data and picking a couple points on this chart to talk about. There were so many more observations predicted that the person was not going to buy the caravan insurance so as that true negative rate increases the true positive rate decreased because it just fills the predictions with more no’s but you can see at 50% TNR the TPR is at 78% so when the TNR is 80% the TPR is 40%. The more people being considered floods the accuracy, meaning that it is rare for a person to want to buy caravan insurance as you can see.

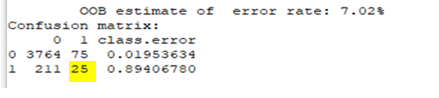
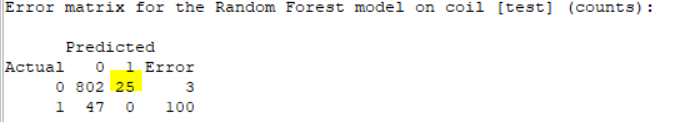
(see chart on next page)



The ROC curve chart with the test data plots the false positive rate (predict yes/ actually no) on the x axis and true positive rate (predict yes/ actual yes) on the Y axis. The plotted 45-degree line represents, on average, the performance of the random variable. The further away the red line is from the diagonal line, the more accurate the model is and as you can see in the chart our red line is not as far from random selection as we would like. We can see that at about 25% False Positive rate the True Positive rate is 50%, the TPR takes a nice jump from 50% to about 75% when the FPR goes from 25% to 50%. From 60% FPR to 100%, the random selection line and our red line don’t separate and the closest both rate’s get to being even is when the FPR is at 80% and the TPR is at 90%. The AUC (area under the curve) is 64% with the test data, dropping 6% from the ROC with the validation data, at 64% AUC, which is ok for accuracy measures.

ROC CHART

The steps I would take to implement a successful continuous feedback method is first, I would look at my scores from the evaluate tab 🡪 clicked on score 🡪then execute the csv file and open it with excel. This gives me a score off all the people who are more likely to buy this insurance, I would give that to sales and marketing with the customers ID’s and scores. Have them create target marketing plans based off those who have a high percentage and see if they would buy. I would then have once a week meetings with management from sales and marketing to go over success rates of the high probability people, then take those scores and adjust my model even more. This way I am taking the results from my model and saying, “ here are the people most likely to buy” and telling sales to call them, if they are unsuccessful, I go back to adjust my model again until we get better results. I scored my model with the test data set, clicking on the class output, and using all the variables. I chose all variables because for the ones that get put in the class of 1 (buying caravan insurance), I want to see what variables were used to determine those people who were put in class 1. What was interesting was that I had 25 true positives from my training data in my confusion matrix, based on my score csv file I had 25 show up with a score of 1, and then comparing that to the confusion matrix that was tested against the test data it showed I had 25 false negatives. That is observations that were predicted no but ended up being yes, so even if it was a True Positive or False negative, the amount of people who bought caravan insurance is still 25. As you can see, with the test data, predicting someone who would buy is at 0% accuracy, which is not good, but what this did was give the insurance company leads to follow, helping them tremendously narrow down people are more likely to buy not guaranteed to buy. The mean of my scores that I ran with probability rather than class is 7% with 25 people having a probability to buy of 50% or higher and of the 47 people that were predicted (see test data confusion matrix) to buy, the mean probability to buy is 10.7%, with none of those percentages being over 50%.



To make this reproducible so others can use my model, I first started with all the defaults for this random forest, then notes my results, checked the error rate chart and see where the line starts to smooth out to first be able to cut down the amount of tree to see how many we can cut down and not give up accuracy. I then will export my log in Rattle that shows the actual code used if someone wanted to use the code in R instead of Rattle. I would say to make this whole process reproducible, it starts from gathering all the documentation starting from how to download the data set from the website into R, adding the heading the variables, the manipulation code that was needed for it, how to get it into Rattle, then using what I said before in starting with defaults, using the log as a reference, and for me using Rattle so code doesn’t have to be typed out every time but with the log present you have the option to use it. Documenting every step of the way from exploring my data, to building the standard tree and explaining why I’m not using it, to my first random forest, to the validation results of my final model. All this information, code, and visuals will be exported and stored in a Dropbox for anyone to look at to try to replicate or expand upon my work. For my pilot plan, I want use a random forest tree and dive into those variables that effected the 12 true 1’s (people who bought caravan insurance), take those variables and target customers who share similar qualities to then create a more focused marketing campaign geared towards the sale of the caravan insurance policy. It starts by documenting those most effective variables, relaying them to sales and marketing teams so they understand the type of qualities that potential buyer of the specific product would have. Take the scores we got from the model which gives each customer a probability of them buying caravan insurance and also target those customers who have the highest scores and then record those results to test the accuracy of those scores. We can target that whole 236 observation’s as well (all people who were predicated to buy but didn’t) as another source of leads for the marketing campaign, you can see already the effectiveness of the model where it helped up narrow down the 236/4025 people who would be more likely to buy (5%) where if you tried to target full amount of people you end up with the problem that arises at this current company. I would then tailor the marketing plan to target that 5%, the people who have high probability scores, and compare the results vs our old marketing results to test improvement in the strategy.

REFERENCES:

* Putten, Peter van der. “The Insurance Company (TIC) Benchmark Original Problem Task Description.” *CoIL Challenge 2000 Report*, 2000, liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/problem.html.
* Williams, G. J. (2013). Data mining with Rattle and R: The art of excavating data for knowledge discovery. New York: Springer.
* Thoko, Justin, and Ted Hessing. “Pilot & Implementation Planning.” *Six Sigma Study Guide*, 22 May 2018, sixsigmastudyguide.com/pilot-implementation-planning/.
* Takaya. “Introduction to the Precision-Recall Plot.” *Classifier Evaluation with Imbalanced Datasets*, 30 June 2017, classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/.
* “Scoring.” *Advantages of the Debian Distribution*, [www.togaware.com/datamining/survivor/Scoring.html#6295](http://www.togaware.com/datamining/survivor/Scoring.html#6295).