Megan Cusey

Week 1 – Caravan Insurance Customer Profile Modeling

02/07/2020

The following details my observations and suggestions for improvement based on the Caravan Insurance Customer Profile Modeling case study taken from Chapter 7 of *Data Mining Applications with R* text. Suggestions are presented following the CRISP methodology: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

### Business Understanding

The text framed its business problem around trying to understand existing caravan policy holders in order to create a profile for them with the objective of creating a more focused marketing campaign towards prospective caravan policy holders. The end result they are hoping for is to increase the sales of caravan policy holders.

The business understanding could be improved. Its worth arguing that most people who believe they require caravan insurance, already have it. The reason people would want to own caravan insurance is that: they own a caravan (or need to insure one on behalf of somebody) that has a value worth insurance, is their primary residence, or is required to as apart of a financing agreement. As a result, the insurance company needs to convince these potential clients to switch from a competitor. Reasons why caravan owners do not insure their caravan: he/she doesn’t have to because the caravan insurance instead required by a financial agency or he/she doesn’t have a loan for the caravan, its value isn’t worth the cost of insurance, he/she can’t afford to, or a combination of these factors.

These hypotheses I would try to confirm with available data. However, I believe if the text’s authors looked at this problem with a different lens, then a better outcome would be the result. Instead, I would look at caravan owners in general (not specifically existing customers) and try to target groups of people who don’t have a caravan now but could in the near future.

### Data Understanding

During data understanding, you want to explore and describe your data set. The text identified that 9822 rows of data exist with 86 features. Of the 86 features, 5 are categorical and 81 are nominal. They discovered that while other religions were represented as binary columns, the Roman Catholic religion was subdivided into a categorical field with 10 levels.

While correlations that exist between the target variable (Caravan Policy Holder) and the dependent variables were investigated, the text seems to have left out the concerns about multicollinearity between other independent variables. Especially because the expected outcome of the case study is to be able to describe caravan policy holders according to variable importance, grouping the potential customers, this is especially important to address. Obvious multicollinearity issues do in fact exist in this data set. For instance, column “APERSAUT”, number of car policies, and “PPERSAUT”, contribution car policies have a correlation of .92. Clearly, as the number of car policies a customer has, the number of contributions made for car policies will also increase.

### Data Preparation

The data preparation step for this case study appeared to be lacking overall. The only things done for this step was 1.) removing the religion variables due to inconsistencies in the way the information was presented and a low correlation with caravan policy holders, and 2.) transforming the categorical variables into binary columns.

Additional steps that I would suggest is to consider some feature engineering steps. Is there a way to combine variables in order to produce a stronger correlation? Is there a way to reduce the number of categorical features by binning them to reduce dimensionality? Also, the issue of multicollinearity between independent variables need to be removed by either combining the variables or removing the (almost) duplicate features.

I would also argue that this dataset might not be the best to address the business issue at hand. The data is static in a nature. I imagine the reason for using this dataset is the availability of information, but ideally some features that might be combined with this information to account for more variation might be their hobbies, number of RV parks where they live, or what they find important in life (through text analytics?). For example, if a customer expresses quality time with his/her family frequently on his/her social media account, then that customer might be more prone to owning a recreational RV to spend time with their family when there’s several different RV parks closer to where they live. It might not be as practical for somebody who lives in a big city where RV parks and space to park an RV is limited.

### Modeling

The models chosen by the case study seemed appropriate. I would have liked them to use the simplest method first to establish a baseline, but they did use logistic regression at the end (even though there were more features selected by the method). I would have added random forest to the mix since they decided not to do much work on their features. Random forest would have “sampled” the different features to give another perspective on feature importance.

### Evaluation

The text evaluated the models according to AUC, recall, and computational performance. It noted that the overall goal of the study was to explain the most about the independent variable (Caravan insurance policy holders) with the least about of dependent variables to reduce complexity. They also sought to obtain interesting results, not those that could be easily explainable without the modeling efforts. For example, those who earn a higher wage and have a boat could be predicted without the use of these methods.

I would be interested to see how these models performed on the test set vs the training set. The variance between the two performance measures is an indication of overfitting. If the test and train set perform similarly, it is an indication that similar performance would be obtained on new data. If the variance is significant, then the model likely has room to reduce overfitting.

### Deployment

The conclusion of the case study indicated that the model built was not ready for deployment. Instead, their objective would be to identify ways to improve the models produced, perform a deeper analysis of the methods and results, and refine the variables supplied to the models. I agree that these steps are necessary before deployment. I would be interested to see results after more feature selection, feature normalization, and model cross validation steps have been performed.