Student Name: Ben Beissner
Student ID: I6336899

Data Mining Lab 03 - Classification

Analytical Report - 09.10.2023

2.3 Assignment 1 Customer Description

in this part of the assignment we want to generate insight about what distinguishes a typical owner of a caravan insurance policy from someone without. Therefore the given dataset will be analysed by using descriptive statistics and white-box models. Visualizations help interpreting the findings and conclusions about typical customers are drawn.

Descriptive Statistics

After loading the data it is important to get familiar with the structure and distribution of the features building the dataset. By accessing the descriptive statistics of the training data we see that we encounter solely numerical features as well as features with numerical categories. Further we find out that we have no missing values in our data. To investigate the dependencies of the features we calculate the correlation of all features and plot this matrix as a heatmap. By doing this we can observe if we have to deal with features that are highly positively or negatively correlated which means that their values rise or fall in the same manner. If this is the case, we can conclude that keeping both variables in our training data will not yield much additional information and therefore dropping these features is beneficial due to dimensionality reduction. The obtained matrix is shown here 1.

One can conclude from looking at the matrix, that there are multiple feature with either high or low correlation. By setting a threshold for the correlation to 0.7 and -0.7, we drop one feature of the strongly correlation feature pairs and achieve dimensionality reduction.

It makes sense to further investigate the distributions of the features. By plotting histograms, we can see how the occurring values of the training data are distributed and can detect if they have high variance or not 2.

We observe that several features in our dataset are represented by only one class. This means they have zero variance and therefore little to now influence in our predictions later. Discarding them is therefore beneficial since we do not loose much information and can further decrease our dimensionality of the training data. After this we are left with 39 features which is a lot easier to handle than the original 86 features of the dataset.

One further important observation about the training data can be made. When looking at the distribution of our labels (CARAVAN POLICY), we see that class 0 (no policy owner) has 5474 out of 5822 samples and is therefore much more represented than the minority class 1 (policy owner). This is important to keep in mind because we are dealing with a binary classification scenario and an imbalanced dataset can lead to high accuracy values of the classifiers even though they perform very poorly when predicting policy owners (class 1). To deal with this issue we implement SMOTE (Synthetic Minority Oversampling Technique). This model synthesizes new examples for the minority class (policy owner) and by doing that balances the dataset. It is to mention that this increases the sample size of the training data and we now have 5475 samples for each class.

White-box models

Since our main task in this part of the assignment is to detect relevant features that describe a typical caravan insurance policy holder accurately, we further focus on selecting the most important features of our data. One efficient way to progress with this to select a specific amount of most important features to train the white-box models. For calculating the importance of each feature we compute the CHI^2 values of each feature and sort the feature accordingly. With the CHI^2 test we compute the expected frequencies under the assumption that variables are independent so that a high score indicates a strong importance of the feature when it comes to predicting the classes of instances. A visualization of the most important features according to the test can be found here 3. After seeing the scores it is reasonable to keep the 10 most important features to fit our classifiers.

To further clarify the influence of the most important features we can visualize the distributions of the features with respect to the occurrence of policy holders with these values. These visualizations of the four most important features can be found in the appendix 4. These visualizations further help to see which values for the important features a typical policy holder has. one commonly used white-box model are decision trees. The tree represents the sequential splitting process of the instances according to their values of the features. The goal is to fit the tree in a way that classes are separated as clearly as possible at the leaf nodes. This makes it possible to see the splitting rules that describe a holder of a policy. The obtained decision tree by training with the most important features is shown here 8. At the leaf nodes we can see how many non-policy and policy holders are classified there. This leads to the conclusion that a typical policy holder can be described as follows: Contribution car policies ¿ 0.5, income ¡ 30.000 ¿ 3.5, Contribution car polices ¡= 4.5 (pure leaf node). Another way to classify a typical policy holder is by splitting according to: Contribution car policies ¿ 0.5, income ¡ 30.000 ¡= 3.5, Customer Subtype ¡= 8.5, Contribution car policies ¿ 5.5.

Interpretation and Conclusion

Since we are viewing the problem of identifying a typical caravan policy holder from a marketing perspective we want to give actionable solutions that are easy to understand. First of all we could identify the most important features of the dataset with regards to predicting if a person is policy holder or not. The 10 most important features according to the CHI^2 test are: 'Contribution car policies', 'Customer Subtype', 'Rented house', 'Income ¡ 30.000', 'Farmer', 'Social class D', 'Contribution moped policies', 'Skilled labourers', 'Unskilled labourers', 'Contribution fire policies'. we further investigated what values of these features occur often for policy holders. by analysing the decision tree we can conclude a potential marketing target for caravan insurance policies are persons that contribute between 0.5 and 4.5 to car policies and have a Income ¡ 30.00 value of more than 3.5 (clear explanations of this values are not given in the appendix of the task.). Further, people with contribution to car policies ¿ 5.5 income ¡ 30.000 $\leq$ 3.5 and customer subtype $\leq$ 8.5 (8 = middle class families) are interesting as well. From looking at the plots of the most important features and the labels we can further suggest to target the marketing to people with customer subtype 8 and 33 (Lower class large families), Contribution car policies value of 0 or 6, Rented house values of less than 4 and Income ¡ 30.000 values of less than 4.

2.3 Assignment 2 Potential Customer Prediction

In this part of the assignment the goal is to predict a set of 800 customers of the test set that contains the most policy holders.

Data Preprocessing

most of the preprocessing already took place in the first part of the assignment. Most important was to split the data to achieve training and test splits and to bring it in a usable format. Further it was very important to deal with imbalance of the dataset. Here, the upsampling with the SMOTE model helped to achieve a balanced dataset.

Feature Selection/Reduction

Also the feature selection and dimensionality reduction was already discribed in the first part. By dropping features with high correlation and low variances nearly half of the original dataset could be discarded without loosing much information. Further the identification of the most important features with the CHI^2 test was a crucial step towards a simpler dataset.

Training Models and their relevance

To investigate which models performs best for predicting if a customer is a policy holder or not a total of five classifiers were trained on the 10 most important features. The models are decision tree, logistic regression, random forest, naive bayes and support vector machine. These models were trained the predictions for the test data were evaluated according to common metrics.

Model Validation

The best performing model out of these five was the random forest classifier. The average scores for precision, recall, f1-score and accuracy are 0.74 each. All the other classifiers achieved lower scores but where in the range of 0.70. The acual values can be found in the notebook and the confusion matrices in the appendix 9.

Customer Selection

After fitting the data to the classifier, the predictions for the 800 most likely policy holders of the test set were made. To achieve this, the best performing classifier (random forest) was used to predict the probabilities of the instances in the test set to be a caravan policy holder. The 800 instances with the highest probabilities were then saved into a list. The calculated probability of the 800th most likely policy holder was about 0.56 and the list of potential policy holders can be found in the notebook.

interpretation and conclusion

During the experiments it could be seen that several classifiers can perform reasonable well on the problem. However, the Random Forest classifier was able to achieve the best scores. To understand why this is plausible we need to understand how the classifier works. The random forest builds several decision trees and chooses the best one in the end. There scores similar or better than the decision tree classifier are to expect. Further, it is possible that because of the ensemble of decision trees, the random forest was better at modelling the imbalances and underlying patterns in the training data.

Academic integrity Declaration

I, Ben Beißner, hereby declare that I have not used large language models or any automated tools for generating the written answers and interpretations in this Analytical Report.

Appendix

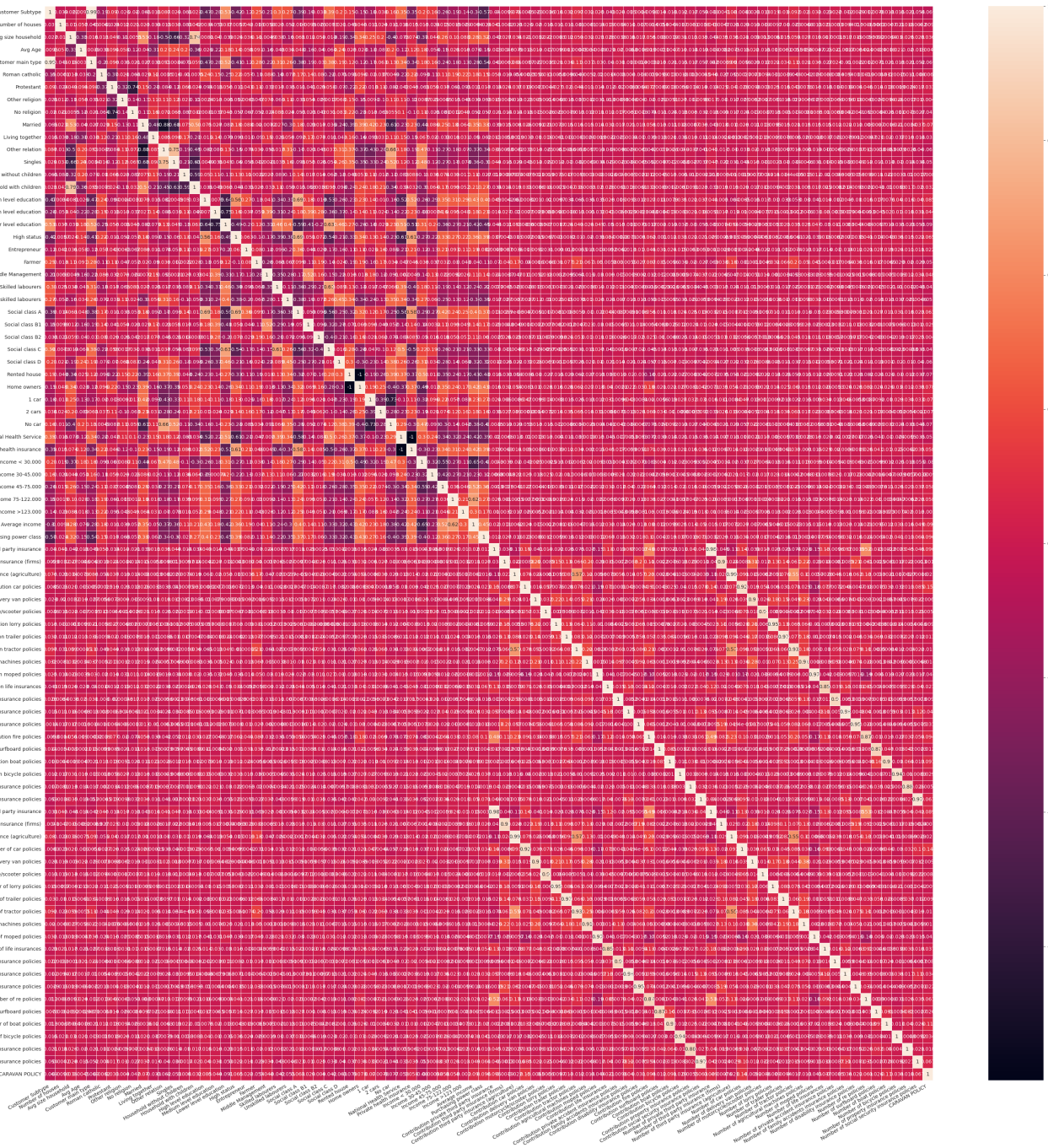

Figure 1: Correlation matrix of training data

Figure 2: Histogram of training data

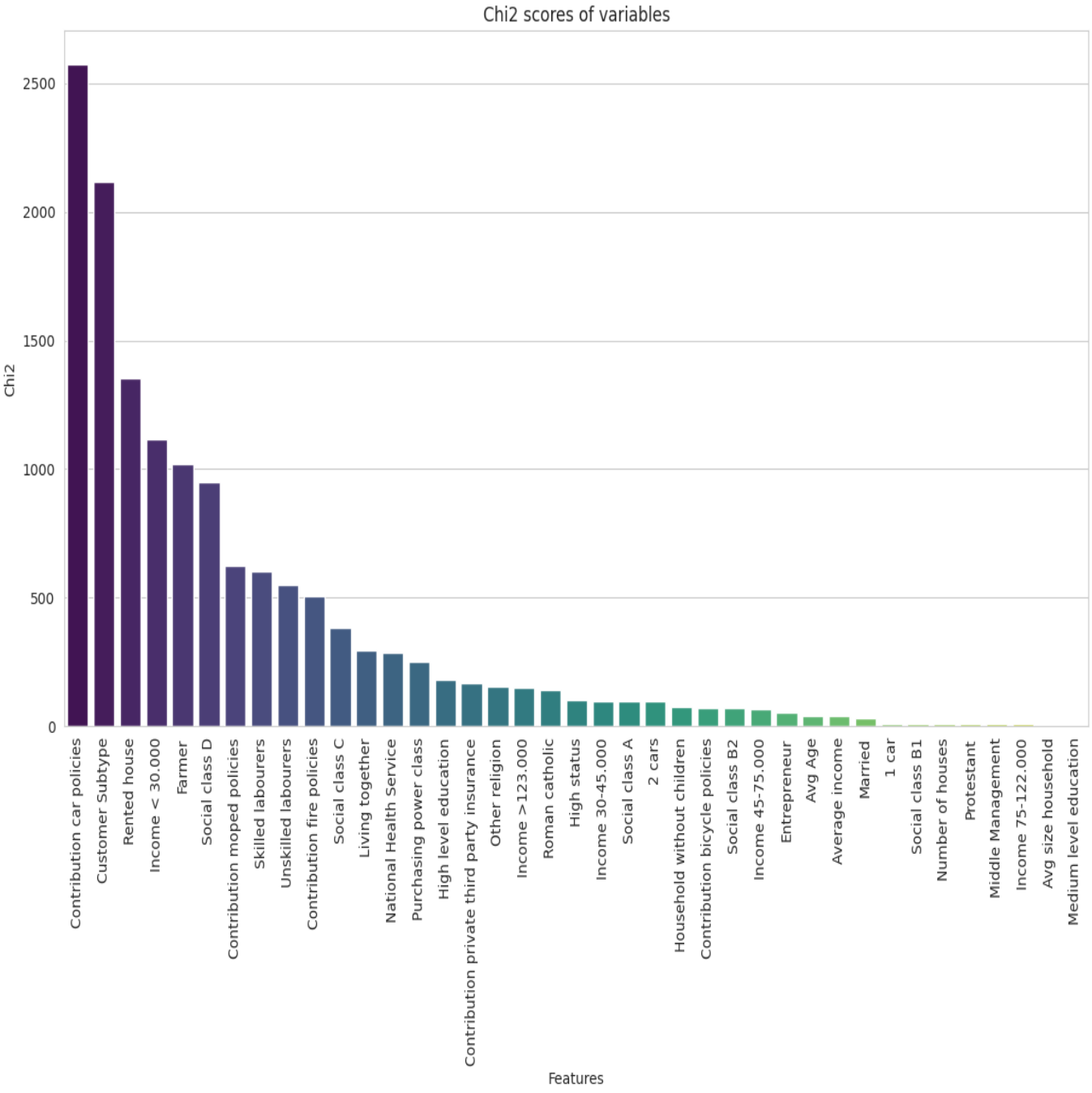

Figure 3: CHI^2 scores of all variables

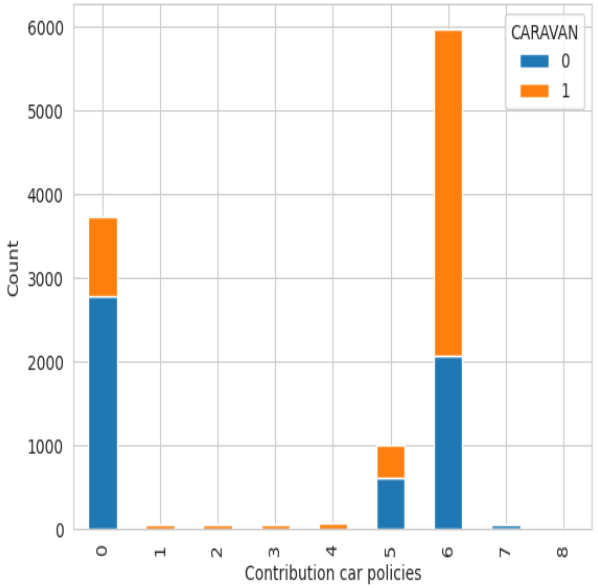

Figure 4: Car policy vs. CARAVAN

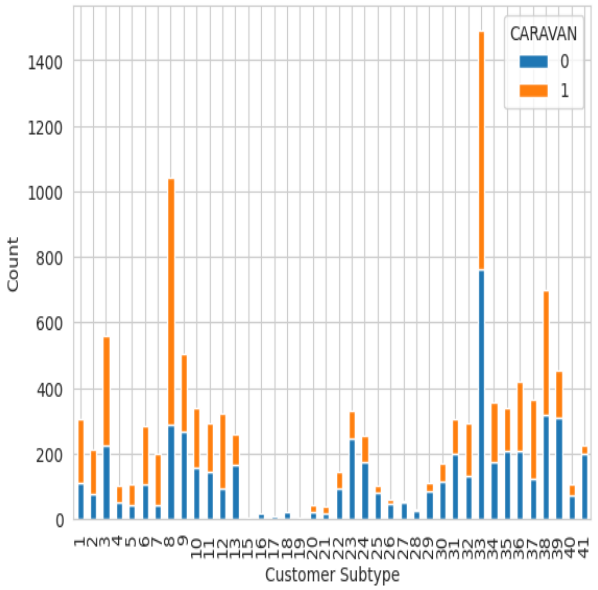

Figure 5: Customer Subtype vs. CARAVAN

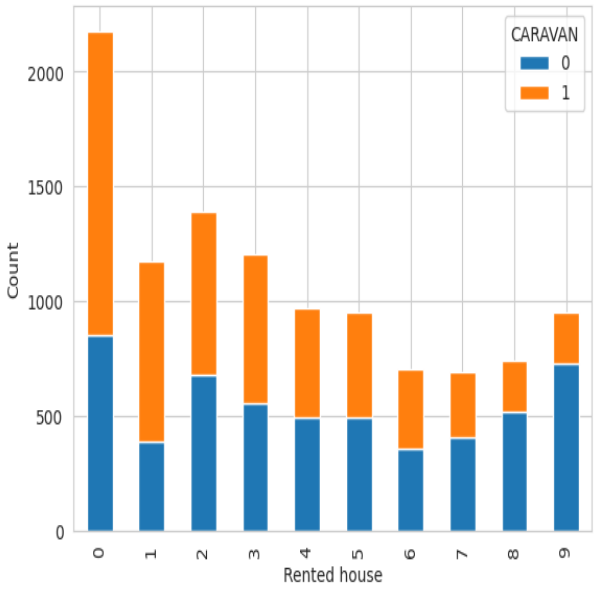

Figure 6: Rented house vs. CARAVAN

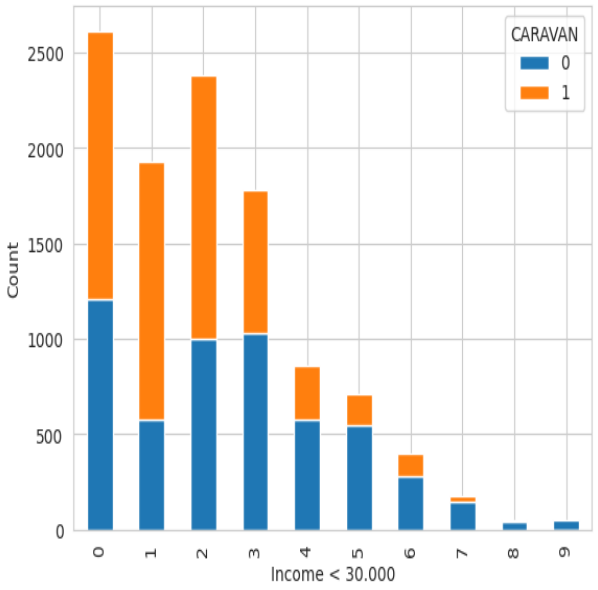

Figure 7: Income ¡ 30.000 vs. CARAVAN

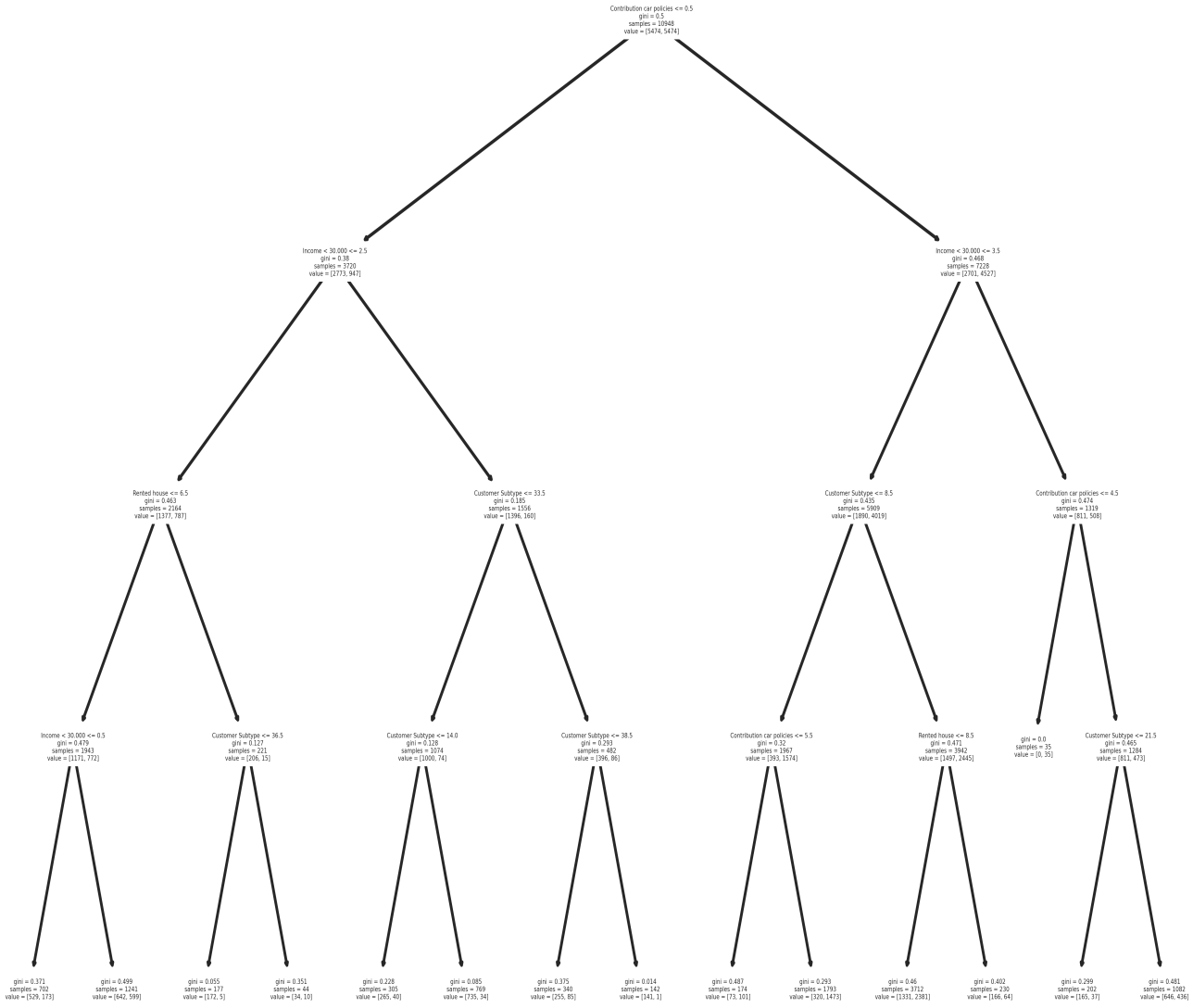

Figure 8: Decision Tree for policy holders

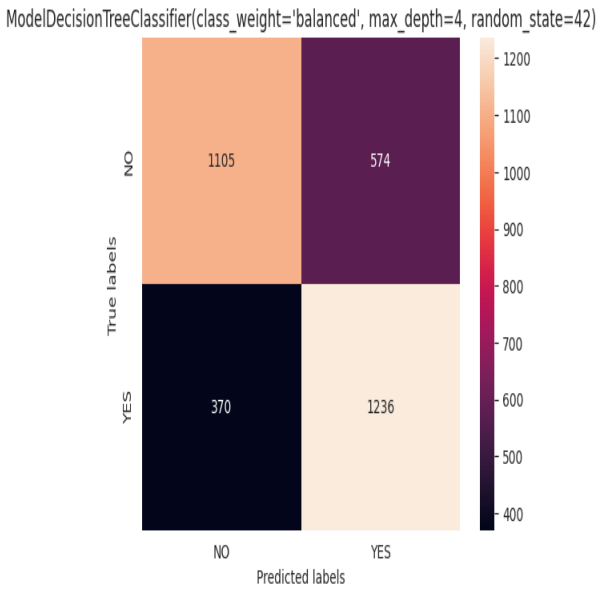

Figure 9: Confusion matrix Decision tree

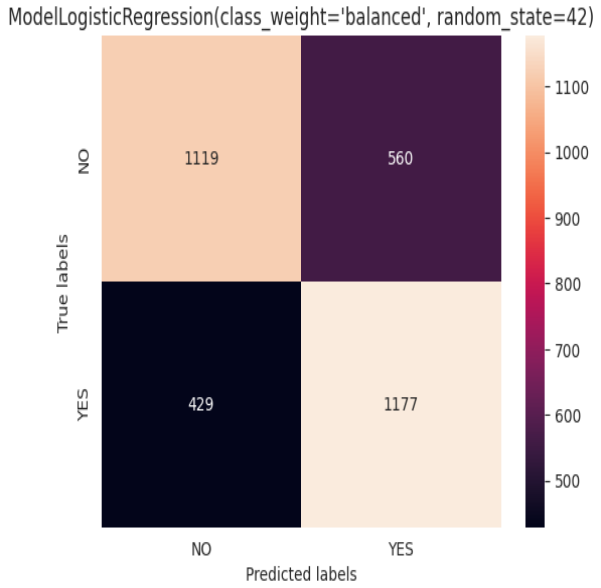

Figure 10: Confusion matrix Logistic Regression

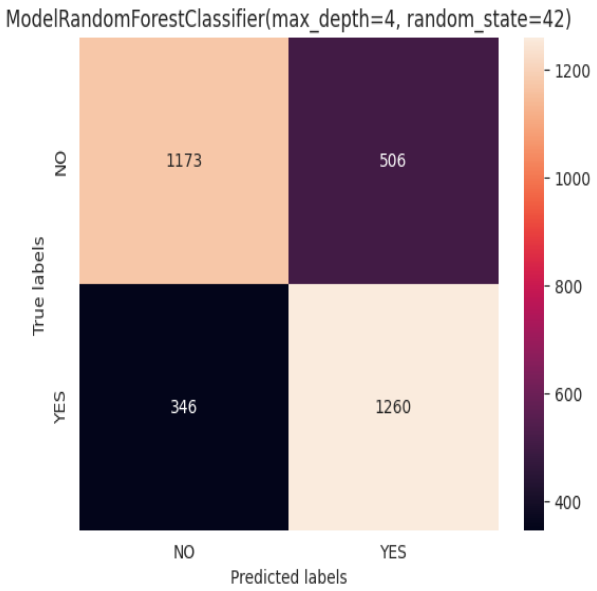

Figure 11: Confusion matrix Random Forest

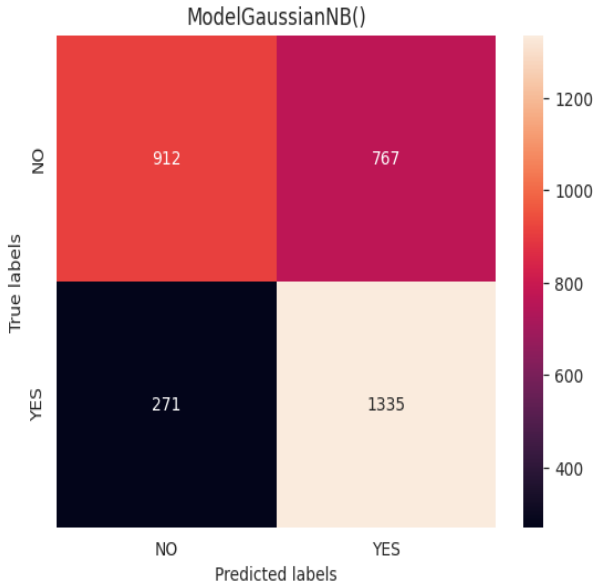

Figure 12: Confusion matrix Naive Bayes

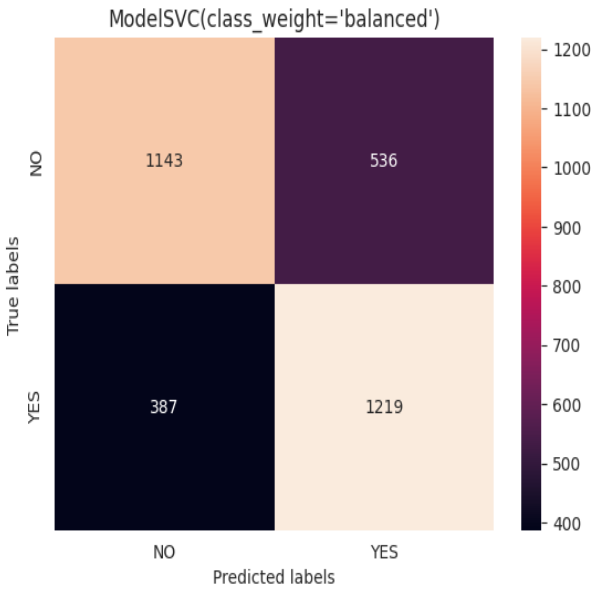

Figure 13: Confusion matrix SVC