11/15/2018

Predicting Loan Default Risk

Predictive Analytics for Business Project

Udacity Nanodegree

Tool: Alteryx

Katerina Bosko, PhD
WWW.CROSS-VALIDATED.COM

The Business Problem

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!

Your manager sees this new influx as a great opportunity and wants you to figure out how to process all of these loan applications within one week.

Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

For this project, you will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to your manager in the next two days.

You have the following information to work with:

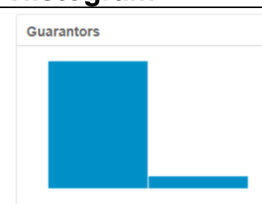
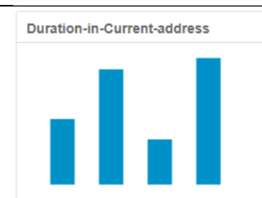
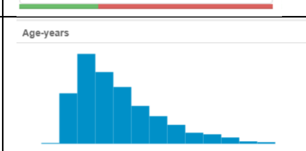
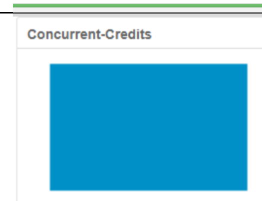
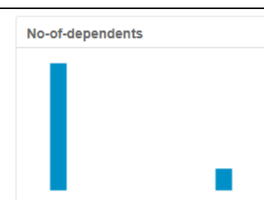
1. Data on all past applications
2. The list of customers that need to be processed in the next few days

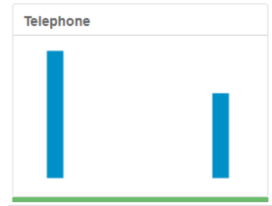
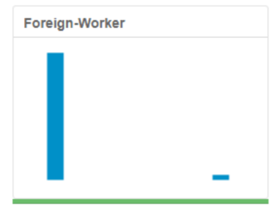
Step 1: Business and Data Understanding

We need to evaluate the creditworthiness of 500 new loan applicants within a week. To do this, we have to create *a binary classification model* based on the data that we have from previous applications. The data we need has to directly relate to the target variable (creditworthiness), e.g. applicant's age, account balance, previous payment history, credit among and its purpose, etc. This will help us understand applicant's ability to pay back a loan.
By applying the trained model on our dataset of new applicants, we will be able to systematically predict which potential customers are creditworthy and which are not.

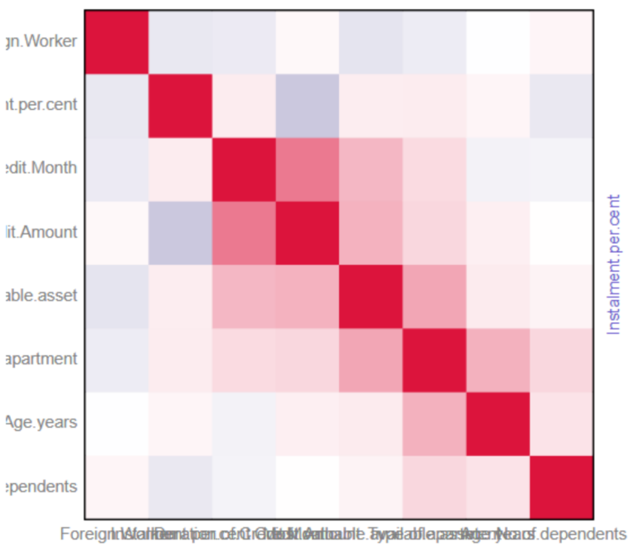
Step 2: Building the Training Set

The following manipulations were done with the original dataset:

#	Field	Status	Reason	Histogram
1	Guarantors	Removed	Low variability (heavily skewed towards one value)	
2	Duration-in-Current-Address	Removed	68.8% missing values	
3	Age Years	Imputed	2.4% missing values imputed with median	
4	Concurrent-Credits	Removed	Low variability (only 1 value)	
5	Occupation	Removed	Low variability (only 1 value)	
6	No-of-Dependents	Removed	Low variability (heavily skewed towards one value)	

7	Telephone	Removed	No logical connection with the target variable;	
8	Foreign-Worker	Removed	Low variability (heavily skewed towards one value); No logical connection with the target variable;	

Using the Association Analysis tool, I created the correlation matrix (see pic below). This showed that none of the variables are highly-correlated (> 0.7) with each other and hence there are no duplicates that could bias the model results.



Pic – The correlation matrix according to the results of the Association Analysis

Step 3: Train your Classification Models

1. ***Logistic Regression:***

Accuracy = 0.7800

Significant Variables (p value < 0.05):

Account Balance, Payment Status, Purpose, Value-Savings-Stocks, Length of current employment, Instalment-per-cent, Most valuable available asset

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

Confusion Matrix (validation set):

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Bias: *more or less balanced compared to other models tested; but still high Type I error*

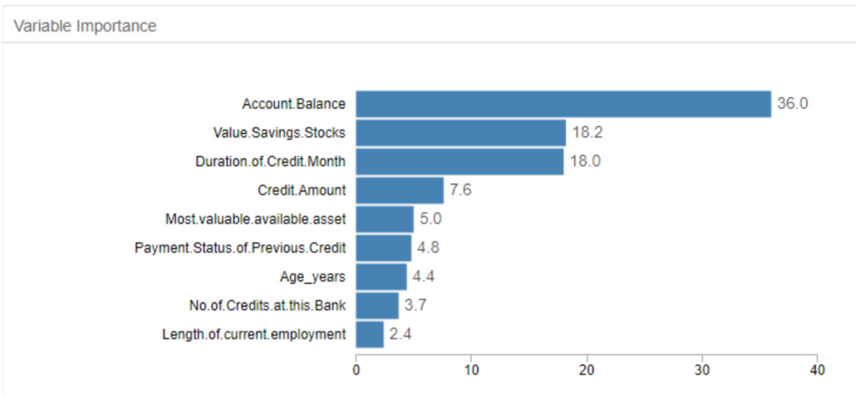
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	TPR* =95/(95+10) = 0.90	FPR = 23/(23+22) = 0.51
Predicted_Non-Creditworthy	FNR = 10/(95+10)= 0.10	TNR = 22/(23+22) = 0.49

** here and later TPR = True Positive Rate; FPR = False Positive Rate; FNR = False Negative Rate; TNR = True Negative Rate*

2. Decision Tree:

Accuracy *= 0.7467*

Decision Tree Variable Importance: Account Balance, Value Savings Stocks, Duration of Credit Month



Decision Tree Confusion Matrix (validation set):

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

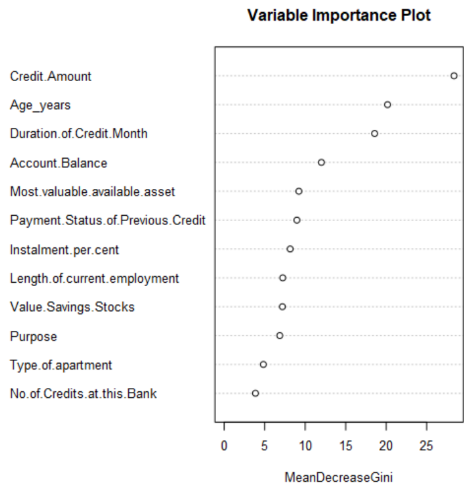
Bias: *high towards false positives*

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	TPR* =91/(91+14) = 0.87	FPR = 24/(24+21) = 0.53
Predicted_Non-Creditworthy	FNR = 14/(91+14)= 0.13	TNR = 21/(24+21) = 0.47

3. Random Forest Model:

Accuracy *= 0.8000*

Random Forest Variable Importance: Credit Amount, Age Years, Duration of Credit Month



Random Forest Confusion Matrix (validation set):

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

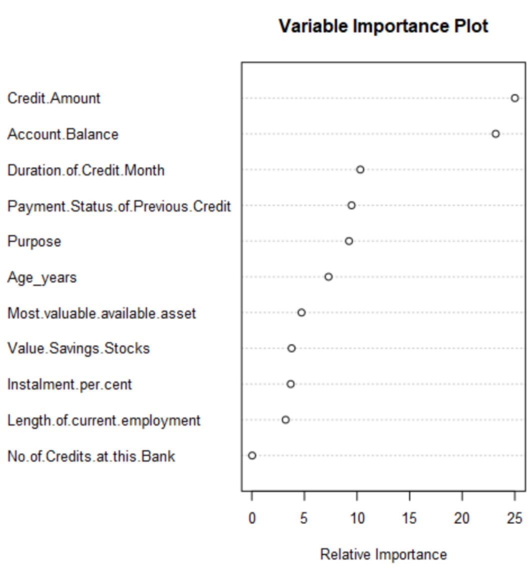
Bias: *high towards false positives*

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	TPR* =101/(101+4) = 0.96	FPR = 26/(26+19) = 0.58
Predicted_Non-Creditworthy	FNR = 4/(101+4)= 0.04	TNR = 19/(26+19) = 0.42

4. Boosted Model:

Accuracy *= 0.7933*

Boosted Model Variable Importance: Credit Amount, Account Balance.



Boosted Model Confusion Matrix (validation set):

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Bias: *high towards false positives*

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	TPR* =101/(101+4) = 0.96	FPR = 27/(27+18) = 0.60
Predicted_Non-Creditworthy	FNR = 4/(101+4)= 0.04	TNR = 18/(27+18) = 0.40

Step 4: Writeup

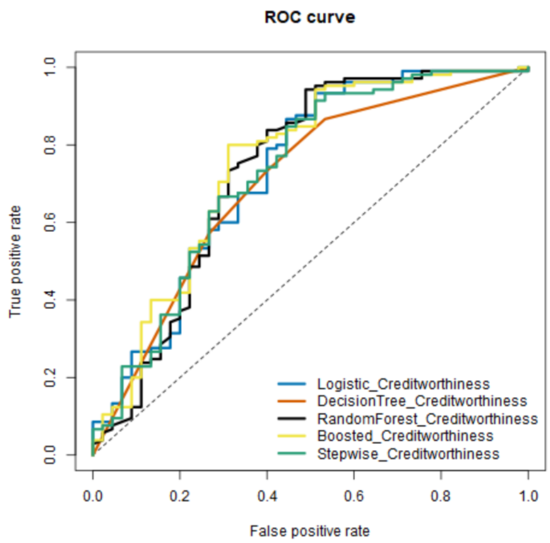
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Creditworthiness	0.7800	0.8520	0.7314	0.9048	0.4889
DecisionTree_Creditworthiness	0.7467	0.8273	0.7054	0.8667	0.4667
RandomForest_Creditworthiness	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_Creditworthiness	0.7933	0.8670	0.7509	0.9619	0.4000
Stepwise_Creditworthiness	0.7600	0.8364	0.7306	0.8762	0.4889

Because my manager cares only about the best possible classification (and not about avoiding the financial losses connected with wrongful classification of uncreditworthy applicants, for instance), I decided to use the RANDOM FOREST model in the scoring part.

Among all models I tried, **the random forest model** seems to perform the best when used on the validation subset. Its overall accuracy is 0.80, with the much better ability to correctly predict creditworthy applicants (true positive rate = 0.9619) than non-creditworthy ones (true negative rate = 0.4222). The random forest model also has the highest F1 score (0.8707) and the second largest AUC (after the boosted model).



As such, the model is optimized for **sensitivity** (biased towards true positives), which is might be ok for a small bank which strategy is to expand, but could be potentially financially detrimental in the long term, since some of the clients are misclassified and have high risk of defaulting (in which case bias towards true negatives makes more sense).

Finally, having applied the forest model to the 500 new customer dataset, I got **406** creditworthy applicants.