INTERNATIONAL BLACK SEA UNIVERSITY

FACULTY OF COMPUTER TECHNOLOGIES AND ENGINEERING

Ph.D. PROGRAM

Data Mining with R Programming Language for Optimizing Credit Scoring in Commercial Bank

Dilmurodzhon Zakirov

Extended Abstract of Doctoral Dissertation in Computer Science

TBILISI-2016

Scientific Supervisor: Nodar Momtselidze

(Professor, Doctor, at International Black Sea University)

(supervisor's signature)

Experts (full name & academic title):

1. **Prof. Dr. Alexander Milnikov**
2. **Assoc. Prof. Dr. Mikheil Rukhaia**
3.

(if any)

Opponents (full name & academic title):

1. **Assoc. Prof. Dr. Abzetdin Adamov**
2. **Prof. Dr. Zurab Bosikashvili**
3. **Assoc. Prof. Dr. Sergo Tsiramua**

(if any)

Introduction

Data mining or Knowledge Discovery in databases (KDD) have already succeeded to attract many research areas to itself. Big Data is being evolved enormously since last two decades. This expansion of data would not mind anything if there were not used data mining techniques and algorithms that have been developed so far. All these techniques and algorithms are used for mining knowledge from huge databases that are growing rapidly.

Databases contain millions of records and there had already been performed lots of researches on these data in order to understand whether there are any hidden facts that are unknown. As these records and data are growing significantly and at the same time growth of database leads to the research and development of new techniques in order to find better results and fulfil current and future needs of large companies that look for such solutions. There exist several areas where this demand is especially felt, such as sales/marketing, buyer behavior, fraud detection, credit scoring *etc*. The progress in keeping and analyzing of data in above mentioned areas leads the organizations with several difficulties while processing and interpreting that big amount of data and turning it into useful information and knowledge.

After enough period of research, it was clear that data mining is that exact approach to meet these challenging requirements. The emerging of this process was directly related with all those hidden patterns and unknown implicit information which lied under that large databases.

Data mining mostly helps to solve problems of Classification, Clustering and Association rules. Since the time data mining has emerged, huge amount of data mining techniques and algorithms have been built in order to extract knowledge from databases. This present work will also concentrate on classification model as majority of techniques and algorithms have been developed for getting out knowledge and information from databases. Classification is a data mining function that assigns items in a collection to target categories or classes. (http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746) The aim of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks which is actually done in our present research. In classification basically used this type of approach: data classes are predefined, a train set of labeled objects are used to form a model through classifier for classification of future observations, also known as test set. For example, a

classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case. This kind of a processing is named as a supervised learning. For solution of classification problems usually supervised learning techniques are used.

Structure of the Thesis

The contents of this thesis are organized in following parts.
First we make Introduction to our research.

The first part is composed of Chapter 1 and Chapter 2.
Chapter 1 is literature review of related work.
Chapter 2 includes detailed exploration of existing supervised learning models.

The second part is composed of Chapter 3 and Chapter 4.
Chapter 3 will be a proof of concept. It will be based on test dataset with exact amount variables as in production dataset but with less amount of data. **This is one of the most important chapters as here we will experiment all of our requirements.** During proof of concept our data source will be comma separated value. The result will affect all of the study.

Chapter 4 will be implementation resulting work in chapter 3. We will use production dataset with enough amount of customer data. This data will be stored in Oracle database. And we will implement only on the most appropriate and accurate model which we will choose from chapter 3. It also includes experimental results and their evaluations resulting from this section. At the end we will estimate predictions and we will look towards optimization of existing credit scoring system, as this is one of the main objectives of this thesis. As a part of optimization task we may re-consider changes in credit processes that may help get more success in loan profitability.

The third part is Conclusion.
It concludes the research work included in this thesis with future directions and suggestions. It also includes appendices and references used throughout the research.

Methodology

The present thesis is an empirical research, including big amount of experimentation work, which is performed for comparing several techniques of supervised learning.

Briefly, we can list the whole study in the following manner:

#	Task Description
	Evaluation of prior researches and works: Research in the area of data mining combined with R programming on credit scoring application is still at a very fresh and preliminary stage. We have investigated and researched existing techniques in the literature. The existing information for present research is reviewed in Chapter 2
	Conception and design of new credit scoring model: Dataset of customer loan information is created which contains all information but will not be used as it is. This will be our production dataset. All dataset is located in Oracle database. So, accordingly data is read from Oracle database and passed to R. Later this dataset will be separated into two types: training dataset and testing dataset. These datasets will be used for prediction purposes.
	Implementation and test: During each model consideration several data activities, such as preparation and cleaning, are being performed. Variables are identified to appropriate classes.
	Evaluation: After evaluation of all considered supervised learning models and estimating best technique we will be implementing this technique on latest and most complete dataset.

	We will compare results of all data mining models chosen at each stage of development so that we will see the quality and accuracy of result. All evaluation is being done on R programming language.
	Dissemination: We distribute the results of this research by submitting to several papers and by reviewing it with experts in analytical areas related with credit scoring thus getting their feedback. The review process will improve further development of this model.

Table 1 Practical approach used in thesis – Research Methodology

Above we described briefly the practical approach of the work done in present thesis. In the upcoming chapters we will be expanding it into details with description of each model and technique and interpretation of the results.

Purpose of the Study

To confirm the objective of thesis, different processes for supervised learning will be analyzed and new developed model shall be proposed with exact results taken from a real life dataset. *Above that all the analyzed models and proposed model will be accompanied by sets of graph plots in order to better visualize the result of some modeling.*

Novelty of the investigation

To propose new algorithms and/or models that will be guided by data mining technology and R programming language and Oracle database to achieve improved and optimized KDD (Knowledge Discovery in Databases) processing from credit scoring system of a bank. **This assumes to be the novelty this study will be bringing.**

Scientific and Practical Importance

We can categorize the contributions of this thesis into following categories:

Contributions to Data Mining

In present thesis different techniques used worldwide for supervised learning were considered and experiments were performed for different models, Decision Trees (Classification and Regression Trees, C5.0, Random Forests, CHAID, Ensemble of Trees), Neural Networks (with multiple hidden layers, SVM and deep learning), Logistic Regression, kNN Classification, Bayesian Classification (Tree Augmented Network), Ensemble of Experts. During experimentation an effort was made to probe different factors related to these models as well as different techniques based on different metrics generated from enormous experimentation.

The outcome shows that different data mining models accompanied with dataset containing several variables can show different accuracy rates. Thus, it can be proposed that variables of dataset and complexity of model used may have different result in the output.

Contributions to Banks and Financial Institutions

As stated in "Problem and Motivation" part of the thesis any financial institution and bank has enormous need in credit scoring system. Purchasing a ready application is always possible but has several disadvantages, such as pricing and licensing, complexity of installation and usage if a financial institution or bank is not so big, support etc. **By building a model with best accuracy rate for prediction it can later be applied to any size of organization for future usage as the variables used in training and test datasets are almost similar for all other financial institutions and banks.** The other important factor to note is, again as stated earlier, that dataset is being stored and read from Oracle database, which is extremely important, as Oracle database is de facto database standard used in banks worldwide. **So, briefly, model can be considered as general purpose and can be used in any banking institution for solving credit scoring issues or optimizing existing one if there is any**.

Structure and volume of the work

The thesis study is 120 pages and consists of four chapters, a list of references, list of figures and list of codes.

Chapter 1 Literature Review

The main objective of this literature review was to gain overall information about data mining, R programming and credit scoring technology. Data mining has many possibilities that can be implemented. R programming is at its beginning stage and will show its power. Credit scoring is also developing year by year.

Chapter 2 Data Mining Techniques and Models Used in The Study

At current time there are dozens of algorithms and techniques which are used for supervised learning. The main problem is to find a suitable algorithm for extracting data for a specific problem case. Same situation applies to credit scoring in financial institutions. This motivated us to research and analyze the factors that affect the selection of an appropriate algorithm of supervised learning and to optimize the overall process of credit scoring by using those methods of data mining in conjunction with programming interfaces. *Credit scoring is one of the key factors for sustainable development of a bank, for increasing of customer amount and for gaining profit.* But in order to have success in this there should be an effective tool and/or mechanism which will consider all possible outcomes of a customer loan application and it's scoring as non-effective scoring will lead to a negative result such as non-performing loans, loss of customer loyalty and of course decrease of income and profit. Although the **bank is not very big in sizes** and credit applications are not so much this does not mean that effective credit scoring is not required. As the number of customers grow and types of loans increase credit scoring is a must for any bank regardless of its size. A system should not be dependent on human factor. *Of course, although decision will be given by a human, not by a software system, a human should have strong predictions based on quantitative and qualitative data gained from customer loan application.* Using standalone software systems will not give that result compared to systems developed on statistics based systems and implemented using analytical tools.

Therefore, it was decided to implement a scoring system which will be modeled on data mining techniques and developed on statistical based programming language.

The present research, first of all, focuses on analyzing different methods of supervised learning. Secondly, after working on existing train and test datasets and applying several methods for these datasets, it proposes a new improved and optimized process for knowledge extraction. During this period huge amount of experiments with datasets is carried out in order to understand which model has which level of accuracy. Thirdly, the present research focuses on developing knowledge and decision based credit scoring system.

Chapter 3 Proof of Concept – Experimentation of Techniques Selected and Evaluation of Them Based On Test Dataset

We use unique approach when modelling; reading data from CSV file, cleaning necessary parts of dataset, dividing ready dataset into training and test data, build model using specific libraries of R language, visualize the results by using different graph plots, use confusion matrix where applicable and at the end interpret the result and decide whether this data mining technique can be used for our model.

First thing we do in all models is read data. For comparative analysis part of all models we are going to read data from CSV file. Next we assign each column data to a specific variable so that we can identify them in model. Later steps of data cleaning and preparation activities follow. To find out the levels in currency variable and inconsistent data is handled here. Two new variables have been created. **"Pldg_Cur"** and **"Pldg_Pledge"** are the two new variables which are derived from **"Pledge"** variable and consist of the number of guarantor and pledge. Purpose of loan consists of many levels and the values have been categorized under a new derived variable **"loanPurpose"**. Hence, there are only few levels in the **loanPurpose** variable. The data related to 'bad insolvency'/'insolvency of borrower' is considered as 1 and others as 0 for **"credithistory"** variable which is derived from "reason of reject or approved".

To make the data available for model creation, it is divided into Training and Testing datasets. Training data to be used to train the model (e.g. data mining model) and Testing dataset to be used to validate the model.

Above steps were common for all model types no matter standalone or ensemble. For visualization of data distribution and accuracy value we use several plot types, in our case we mostly use ROC

curves, as they are specifically developed for such purposes. Example of such ROC curve for decision tree data mining technique can be seen below.

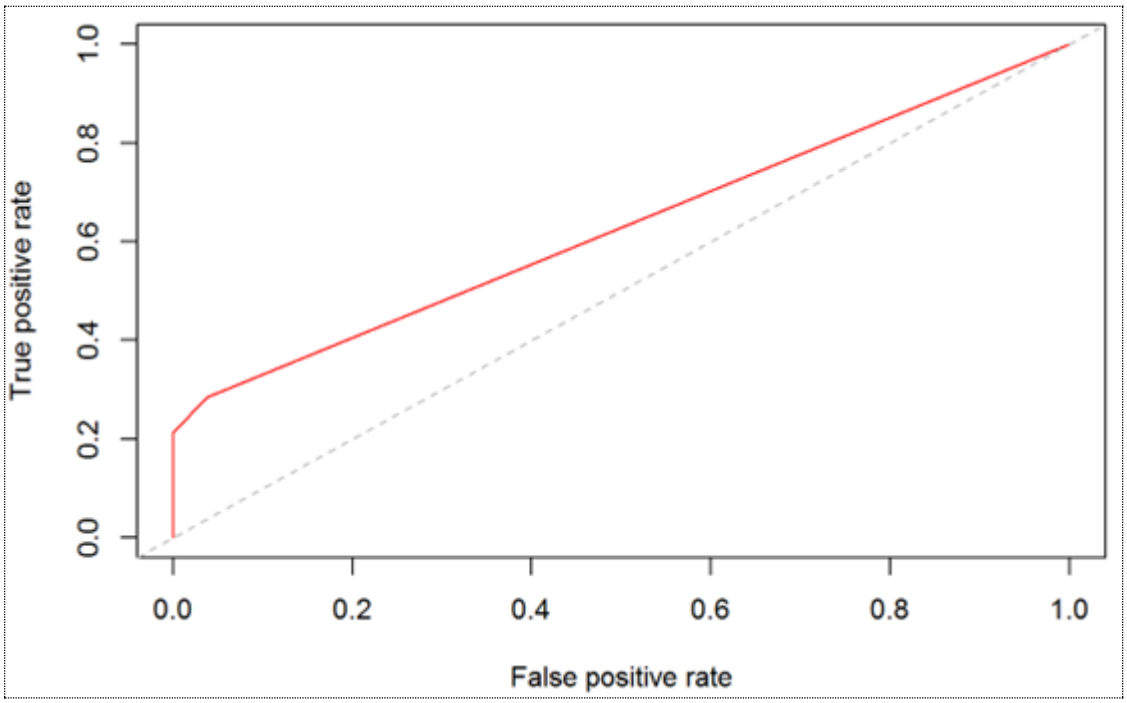


Figure 1 Decision Tree test dataset ROC curve

As the system is a kind of machine learning type thus we have to teach our model first and then it will calculate rest data by itself. In our case we use CSV file data for training and testing our model and after getting satisfied results we will use Oracle database data for validation. So in this chapter along with data source we have new additional process for our modeling, validation process. Validation is like production environment; all training and tests are done on external file and validation is done on real database data. We chose Oracle database not by chance. It is worldwide practice that Oracle database is used in most of mission critical transaction systems. Such systems are mostly implemented in financial, telco and banking sectors. Combination of such complex and high-end systems like machine learning and making data source Oracle database will surely outcome in a powerful solution for deep learning system.

Shortly training and testing process will be as follows:

- We train our model based on both Random Forest and Random Forest UnderSampled data mining technique.
- Later we build an ensemble model consisting of data mining techniques from Chapter 4.
- As a last step we compare Random Forest models and ensemble model in order to determine which has best accuracy and fits our model best.

Next we build two models; Random Forest model and Random Forest UnderSampled model.

Now we will create a function which provide methods for collection, analyzing and visualizing a set of resampling results from a common data set.

The summary of function computes summary statistics across each model/metric combination. From same dataset, it resampled and applied all models in list**. And generated statistics for accuracy, such as what was the average accuracy found for all 10 samples per model.**

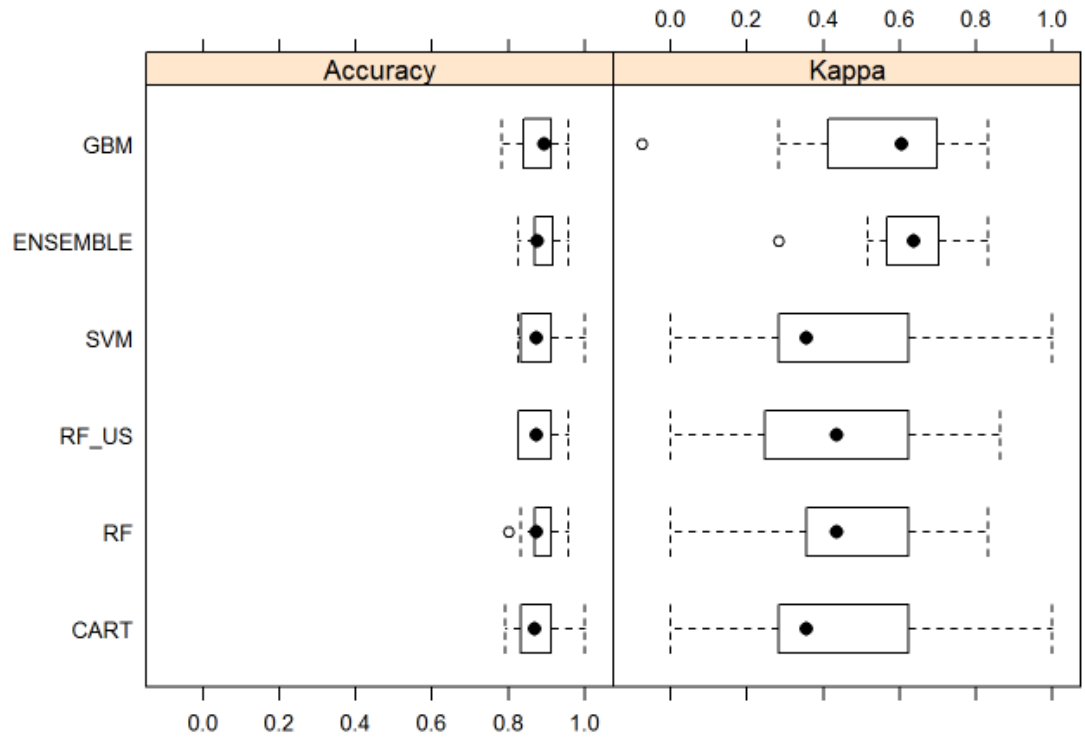
```
results <-
resamples(list(ENSEMBLE=ensemble_experts,RF=rf_model,RF_US=rf_model_US,CART=c
art_model,GBM=gbmFit2, SVM=svm_model))
summary(results)

## Models: ENSEMBLE, RF, RF_US, CART, GBM, SVM
## Number of resamples: 10
##
## Accuracy
##             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## ENSEMBLE 0.8261  0.8696 0.8748 0.8891  0.9158 0.9565    0
## RF       0.8000  0.8696 0.8723 0.8813  0.9130 0.9565    0
## RF_US    0.8261  0.8279 0.8723 0.8757  0.9130 0.9583    0
## CART     0.7917  0.8350 0.8696 0.8813  0.9130 1.0000    0
## GBM      0.7826  0.8474 0.8940 0.8853  0.9130 0.9565    0
## SVM      0.8261  0.8424 0.8723 0.8808  0.9035 1.0000    0
##
## Kappa
##                Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## ENSEMBLE  0.28360  0.5712 0.6363 0.6243  0.6902 0.8321    0
## RF        0.00000  0.3551 0.4343 0.4563  0.6230 0.8321    0
## RF_US     0.00000  0.2553 0.4343 0.4133  0.6230 0.8636    0
## CART      0.00000  0.2841 0.3551 0.4358  0.6230 1.0000    0
## GBM      -0.07477  0.4372 0.6049 0.5328  0.6788 0.8321    0
## SVM       0.00000  0.3015 0.3561 0.4394  0.5956 1.0000    0
```

Figure 2 Function for methods of collection, analyzing and visualizing results from a dataset

Now let us see how these results are fit in box-and-whisker plot and in dotplot. Resampling is validating models by using random subsets, i.e. cross validation.

bwplot visualizes the results of accuracy after resampling. A good model should exhibit very less variance in accuracy across all samples. We can see that Random Forest (RF) has a mild outlier model – not good result. RF_US has no lower bound. Probably with least variance. Check out the variance in CART – quite large.

```
bwplot(results)
```




```
dotplot(results)
```

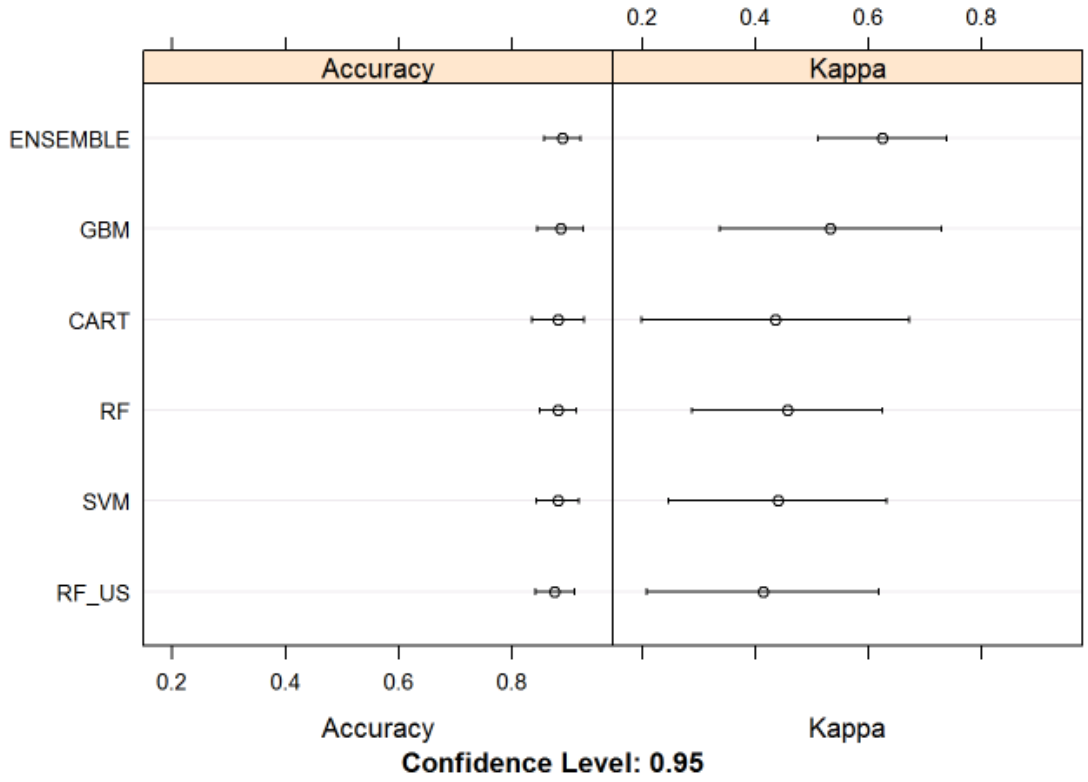


Figure 3 Visualizing Function for methods of collection, analyzing and visualizing results from a dataset

And now we can look for the differences between models above by using function we created. This graph makes pairwise comparison of models. It is not very insightful in our case but can be used if we had two very close cases.

Kappa is a measure of how well the classifier performed as compared to how well it would have performed simply by chance. In other words, a model will have a high Kappa score if there is a big difference between accuracy and the error rate.

```
difValues<-diff(results)

summary(difValues)
## Accuracy
##           ENSEMBLE RF          RF_US       CART        GBM         SVM
## ENSEMBLE           7.819e-03  1.336e-02  7.804e-03  3.819e-03  8.290e-03
## RF        1                   5.543e-03 -1.449e-05 -4.000e-03  4.710e-04
## RF_US     1         1                   -5.558e-03 -9.543e-03 -5.072e-03
## CART      1         1         1                    -3.986e-03  4.855e-04
## GBM       1         1         1         1                      4.471e-03
## SVM       1         1         1         1          1
##
## Kappa
##           ENSEMBLE RF         RF_US      CART       GBM        SVM
## ENSEMBLE            0.167947  0.211028  0.188536  0.091453  0.184895
## RF        1.0000               0.043081  0.020589 -0.076494  0.016948
## RF_US     1.0000    1.0000               -0.022492 -0.119575 -0.026133
## CART      1.0000    1.0000    1.0000               -0.097083 -0.003641
## GBM       1.0000    1.0000    1.0000    1.0000                0.093442
## SVM       0.9654    1.0000    1.0000    1.0000    1.0000

bwplot(difValues, layout = c(3, 1))
```

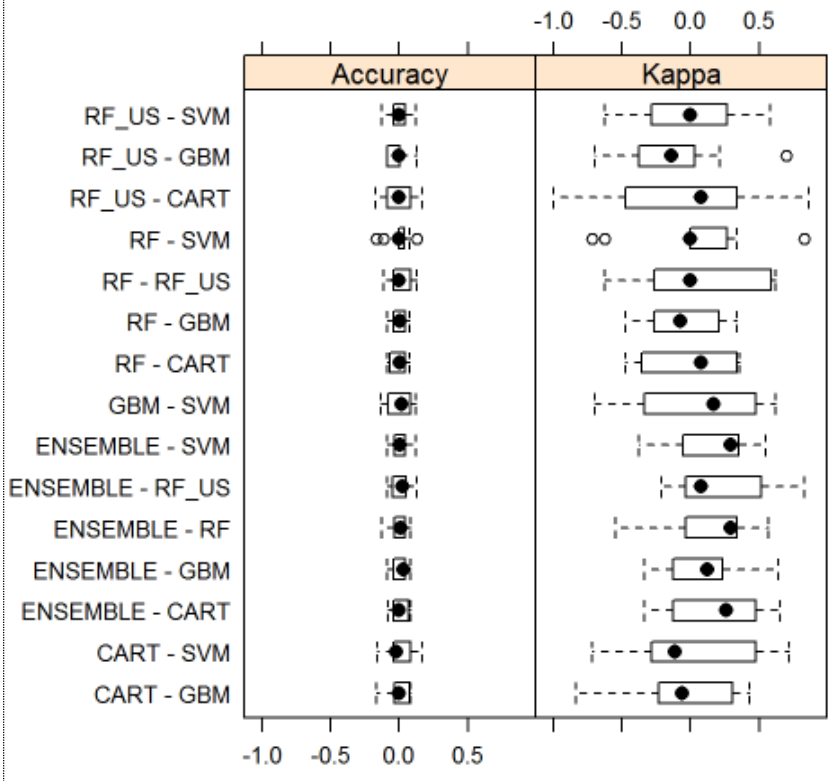


Figure 4 Indicating difference between models

Chapter 4 Implementation of Final Model Developed On Production Dataset and Its Results, Interpretation and Evaluation

Up to now we built our models and made an analysis on how they perform well compared to each other. And now we have to make calculation in order to decide which of the models we built is the best to be used in our research.

In order to do that we created a function valAtRisk. **This is a function that calculates the classifiers based on least amount of value put at risk.**

```
valAtRisk<- function(cross_tab_obj){
  saved_money<-1200*cross_tab_obj[1]
  val_risk<-70000*cross_tab_obj[3]
  xtra_revision<-1200*(cross_tab_obj[2]+cross_tab_obj[4])
  total_var<-xtra_revision+val_risk-saved_money
}
```

Figure 5 Function for calculating Value At Risk amount

Now that we have a function, we need to apply it to each of our models we built and according to the output value we will make a decision. **And in the output we have RF_US model.**

```
rf_var<-valAtRisk(rf_ct)
rf_US_var<-valAtRisk(rf_US_ct)
gbm_var<-valAtRisk(gbm_ct)
svm_var<-valAtRisk(svm_ct)
cart_var<-valAtRisk(cart_ct)
nb_var<-valAtRisk(nb_ct)
knn_var<-valAtRisk(knn_ct)
ensemble_var<-valAtRisk(ensemble_ct)
var_frame<-
data.frame("Model_Name"=c("RF","RF_US","GBM","SVM","CART","NB","KNN","ENSEMBL
E"),"VaR"=c(rf_var,rf_US_var,gbm_var,svm_var,cart_var,nb_var,knn_var,ensemble
_var))

best_model<-var_frame[which.min(var_frame$VaR),1]

print(paste0("The best model is ",best_model))

## [1] "The best model is RF_US"
```

Figure 6 Values at Risk Calculation

For better understanding let us give a confusion matrix output for both Random Forest models and try to clarify them. We already know that confusion matrix is a table that is used to describe performance of a classification model on a set of test data for which true values are known.

```
print(rf_model$finalModel)

## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 11.97%
## Confusion matrix:
##           Approved Rejected class.error
## Approved      191        0   0.0000000
## Rejected       28       15   0.6511628


print(rf_model_US$finalModel)

## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 21.79%
## Confusion matrix:
##           Approved Rejected class.error
## Approved      159       32   0.1675393
## Rejected       19       24   0.4418605
```

Figure 7 Confusion matrix for Random Forest models

rf_model is a list of many random forest. Rf_model$finalModel is the best random forest model formed. We use a training framework where in one command we train many models in one go. The output of rf_model/rf_model_US will list the accuracy of all types of models built and rf_model_US$finalModel will give you the particulars of best model. It will also give the confusion matrix of training data.

In confusion matrix; row is the actual application status. Column is predicted from model. So out of 178 approved applications, 159 were correctly identified as approved by the model. Similarly, out of 56 rejected applications, only 24 correctly identified by the model.

class.error is the error rate in identifying correct class, that is accuracy. In other words, how often is the classifier correct.

Conclusions

In this research study we started with analyzing data mining techniques. Literature review of mostly used data mining algorithms have been performed and several examples were reviewed. We talked about what data mining and machine learning does and what issues they still have as a missing. We listed tasks of data mining, such as classification, regression and etc. As our primary study relies on finance sector we also reviewed shortly application of data mining in finance sector.

Next we reviewed our tool which we used for development, which is R programming language. We compared it with other competitors in the market and listed its capabilities and strengths. We saw that R language is a strong player in the market of statistics, data mining and machine learning area, despite the fact that it is relatively new programming language.

As a last part in literature review we talked about credit scoring, our main research topic. We talked about its history and development stage and how it is important to have a credit scoring for a financial organization. We discussed the credit scoring is made, what parameters and variables are important during processing credit scoring.

After literature review we deeply analyzed each of the data mining technique which we planned to experiment in future sections. We described ways they perform data mining and machine learning.

After finishing analysis of data mining techniques in theory, we had to see how each of them perform in practical experiments. Since our research aim is to optimize a scoring we had to first understand the work of each algorithm and then choose best performing technique and then use and tune it for optimizing a scoring. In order to be fair, during proof-of-concept we used same dataset in every data mining algorithm. During all experiments we used R packages for data mining and datasets for training and testing of models.

So as we have made comparative analysis of all selected algorithms we had to choose best performing technique to build our model and validate our data. We chose Random Forest UnderSampled algorithm for building our credit scoring model. In last two chapters we used code fragments in order to better understanding the output result and used several graph plots for visualizing of training and tests.

As a conclusion we would like to mention one important aspect; at the end of research we chose best model but that did not mean that other models are bad. They are not bad, but for the time being they are just not as good as our best model. Data mining and machine learning are deep sciences and there is no excellent model or techniques to solve a specific task. It all depends on what task are you performing on specific data mining algorithm and what is your dataset. Not having standardized dataset will leave us with the risk of having unsatisfied results. In our case the reason why we chose Random Forest UnderSampled is that by that

way best financial value for bank is delivered, in other words using this technique means bank loses least amount of money.

Up to now our computations show that UnderSampled Random Forest model is best in predicting and modelling our dataset.

The reason is that there is a cost associated with wrongly classified rejected as accepted, and at the end the bank or any other financial institution will have to take the risk.

So the model that minimizes that risk and delivers best financial value in our study tends to be UnderSampled Random Forest and we count it as best and final model.
Most existing classification methods do not work well when dataset is imbalanced. They do not practice sampling methods. In our case we used Random Forest algorithm and used under Sampling to eliminate imbalance on dataset and result is better than other data mining classification methods.

Recommendations

For the future work this research can be extended in several ways;

- The same technique can be used to build model for different task of any other area except than financial.
- Deeper ETL process can be applied in order to have cleaner data.

Publications

1. Zakirov, D. and Momtselidze, N. (2015) Application of Data Mining in the Banking Sector, *IBSU Journal of Technologies and Technical Science*, 4(1), p.13-16, ISSN:2298-0032.
2. Zakirov, D., Bondarev, A., and Momtselidze N. (2015). A Comparison of Data Mining Techniques in Evaluating Retail Credit Scoring Using R Programming, *12th International Conference on Electronics Computer and Computation (ICECCO) IEEE*, p.69-73, ISBN: 978-1-5090-0199-6
3. Zakirov, D., Bondarev, A., and Momtselidze N., (2015). Data Warehouse on Hadoop Platform for Decision Support Systems in Education, *12th International Conference on Electronics Computer and Computation (ICECCO) IEEE*, p.73-77, ISBN: 978-1-5090-0199-6

4. Zakirov, D. (2016) Credit Scoring based on Random Forests Algorithm: An Effective Empirical example, *Journal of Science, Innovation and New Technologies of Kyrgyzstan*, 1(2016), p.44-49, ISSN:1694-7649.