Customer Churn Prediction for Retail Business

Annapurna P Patil
Dept. of Computer Science & Engineering
Ramaiah Institute of Technology,
Bangalore, India
annapurnap@msrit.edu

Savita Shetty
Dept. of Information Science & Engineering
Ramaiah Institute of Technology,
Bangalore, India
savita_ks1@msrit.edu

Deepshika M P
Dept. of Computer Science & Engineering
Ramaiah Institute of Technology,
Bangalore, India
deepshikamp@gmail.com

Samarth S Hiremath
Dept. of Computer Science & Engineering
Ramaiah Institute of Technology,
Bangalore, India
samarthhiremath28@gmail.com

Shantam Mittal
Dept. of Computer Science & Engineering
Ramaiah Institute of Technology,
Bangalore, India
shantammittal@gmail.com

Yogesh E Patil
Dept. of Computer Science & Engineering
Ramaiah Institute of Technology,
Bangalore, India
patilyogish@gmail.com

***Abstract*— Customer churn happens when a customer discontinues his or her interaction with a company. In retail business, a customer is treated to be churned once his/her transactions outdate a particular amount of time. Once a customer becomes a churn, the loss incurred by the company is not just the lost revenue due to the lost customer but also the costs involved in additional marketing in order to attract new customer. Reducing customer churn is a key business goal of every online business.**

In this project, we have considered data set from UCI Machine Learning repository. This dataset contains records of transactions that occurred between December 1, 2010 and December 1, 2011. This is recorded from an online retail gift store based in United Kingdom. The customers of this company are mainly wholesalers. This data set is pre-processed by removing NAs, validating numerical values, removing erroneous data points. We then perform aggregations on the data to generate invoice based and customer based data sets. A variable churn is attached to each data point. This churn value is determined based on the customer's transactions. Three algorithms are run on this customer aggregated dataset to predict churn value. They are Random forest, Support vector machines and Extreme gradient boosting. A comparative study is done on these three algorithms.

Keywords— Churn, Random Forest, Support Vector Machine, Extreme Gradient Boosting

I. INTRODUCTION

In a retail sector business, some customers stick around while others stop shopping at a particular store after certain period of time. Detecting which customers have decided to shop elsewhere and which ones are idle at the moment, is a Herculean's task for a company. Customer churn is the tendency of customers to stop purchasing with a company over a time period. Customer churn is also called customer attrition or customer defection. Churning impedes growth. Therefore, companies should have a proper defined method to compute customer churn rate for a given time. By keeping track of churn rate, organizations can be equipped to success rate in terms of customer retention.

A. Objectives

The main objectives of this project are listed below:

- To predict churn value for all the customers of the company for a given period of time.
- To compute the overall churn rate for the given time.
- To provide a deeper insight into the sales by analyzing customers' buying pattern.
- To detect customers who are about to drop out from the business in order to take necessary steps.
- To provide clear visualizations of the churn predictions to help businesses come up with better strategies.
- To help businesses know the real value of a potential churn customer and retain him/her as a loyal customer by establishing priorities, optimizing resources, putting efficient business efforts and maximizing the value of the portfolio of the customer.
- To help businesses come up with personalized customer retention plans to reduce the churn rate

B. Deliverables

The product will be a tool that helps in predicting if a customer will remain loyal or not in a given period of time.

978-1-5386-1887-5/17/$31.00 ©2017 IEEE

This, as already mentioned, helps in coming up with proper business strategies and hence, reduce the risk of a particular customer turning into a churn.

The product deliverables for our project are as follows:

- A system which can predict if a customer is a churn or not for a retail business.
- A system which can compute churn rate of the retail business.
- A system which can run multiple algorithms and compare performance among them. Algorithms include Random Forests, SVM and Gradient Boosting to predict the customer churn for a given period of time.

Customer churning happens when a customer chooses to terminate association with a company. Online retail businesses consider a customer as a churn if he/she has not interacted with a site or service for a particular amount of time. When a customer becomes a churn, it generates a loss in revenue due to that customer and the marketing costs involved in order to replace churned customer with a new one. The churned customer's spending to date may not have covered the costs of initially acquiring that customer, apart from the direct loss of revenue. It is a more dreary and expensive task to acquire a new customer than to retain an old and already loyal customer. Hence, reducing customer churn is one of the key business goals of every retail business.

Churn prediction techniques tend to model and understand customer behaviours precisely and record the attributes which denote the risk and time of a customer churn. Since there is a lot at stake when a customer becomes a churn, accuracy in predicting churn is critical for a company's success. By predicting a churn customer, the marketing department of the company can focus on such customer and put proactive retention efforts in order to retain the customer. In addition to this, some retention-focussed incentives and offers may be provided to such churn customers to make them happy, active and satisfied customers.

II. Literature Survey

A. Sequential Patterns

The work on sequential patterns [1] emphasized on a new data mining algorithm, called DEML. The accuracy of predictions made by this algorithm are estimated for handling classification problems. An evolutionary approach was used in DEML searches in performing some task keeping in mind the possible rule space. At first, a probabilistic induction technique was used in order to generate rules, and based on these rules, higher order rules were obtained in an iterative manner.

When identifying interesting rules, an objective interestingness measure was used. The probability that attribute values of record will be correctly determined using rules it encodes was known to be fitness of chromosome. And the subscribers were ranked according to how likely they're to churn based on predictions made. In particular, when applied to telecom subscriber data, it accurately predicted churn under different churn rates.

B. Genetic Modelling

Reference [2] emphasized on results aimed at application and assessment of techniques of data analysis in the field of sales marketing. The applied techniques were: genetic programming, CHAID, rough data analysis and logistic regression analysis. All the four techniques were applied to the customer retention modelling problem independently. For this, a financial company database was used.

The techniques were used to create models which were used to gain insights into making predictions on the relationship ending with the company and influencing factors for customer behaviour. Upon comparing the prediction power of the models, it was shown that the genetic expertise offers the maximum performance.

C. Neural Networks

Reference [3] used predictors by selecting a model class neural network, a subscriber representation (sophisticated and naive), and model combination techniques such as none, averaging, or boosting. For every predictor, they performed a ten-fold cross validation study, utilizing similar splits across every predictor. In each splitting of the data, the ratio of churn to no-churn examples in the training and validation sets was the same as in the overall data set. For the neural net model, the input variables were reformed by subtracting the means and scaled by dividing by their standard deviation. Input values were supposed to lie in the range. Networks were being trained until they were a local minimum in error.

D. Markov Chain, Logistic Regression and Random Forests

The paper under [4] and [5] focussed on the importance of detection of probable potential churners in the earlier stages which aid enterprises in targeting such valuable customers by the use of certain set of actions that are helpful in customer retention and increased profits.

E. Game Theory

Game theory deals with understanding strategic situations where, how well a customer performs, depends on what other customers do and vice-versa. The basic principle of game theory is to find out an optimal solution for a given situation. Game theory is not just used in games like Chess and Football, but also in many other areas where important decisions are made, like investing and customer management.

It is used to study situations that are competitive in a more structured way. In today's world, big data plays a major role in determining success of a business and application of game theory to the data is a smart move that can aid businesses to predict probable outcomes of a business and the interest shown by the customers. Hence, game theory can be applied to predict how rational customers will make decisions that help them make effective savings.

Game theory is applied to solve problems together with machine learning and AI. Customer retention is an interactive contextual problem involving customers as well as enterprises. It involves different levels of engagement, conflict and

cooperation between customers and company. Such problems can be tackled by Game theory and also provide mutually beneficial outcome.

Rather than using continuous variables, Game theory considers discrete variables determining events, actions or outcomes for prediction. It makes assumptions that the engagement between customers and business involves unbiased decision makers and has deterministic outcomes.

Some real-time applications where Game theory predictions have been successfully tested are mentioned here. Nate Silver, a famous New York Times blogger and a statistician, used strategies of Game theory and predictive analytics to predict that President Barack Obama would be re-elected. This has brought combination of Game theory and Predictive Analytics under limelight. In another scenario, a patented Game theory based algorithm was developed by Armorway. This uses big data to show visualizations and develop strategies that are intelligence driven. Currently, this algorithm is being used by University of Southern California in order to improve campus security. It categorizes and classifies different types of vulnerabilities that might occur in the campus. This same algorithm has also been implemented to improvise the efficiency of US Coastguard patrolling. Records show that there has been an increase by 60% in terms of effectiveness of patrolling.

Although Game theory is not a core part of Data Science and Predictive Analytics, it helps enterprises leverage customer behaviour with different set of approaches. Therefore, the prediction of behavioural models by Game theory can deliver rich insights for a company. Using Game theory along with predictive analytics, scientists will soon be able to predict far more accurately the future events.

III. Design

Design is significant phase in development of software. It is basically a creative procedure, which includes the description of the system organization, establishes that it satisfies the functional and non-functional system requirements. Larger systems divided down into smaller sub-systems contain services that are related to each other. The output in design phase describes the architecture of software to be used for the development of the common endpoint service. This section depicts the issues that are required to be covered or resolved prior to the attempt to instrument a complete solution. The detailed design includes an explanation for all the modules. It throws light on the purpose, functionality, input and output.

A. Architecture Design

The system architecture of our project is shown in the figure below. RStudio is an IDE which is where the entire project and its workspace is set up. R is a language and environment for computing in statistics and graphics. It was designed under GNU project. It is based on the lines of S language developed at Bell Laboratories.

This engine has packages underlying to it, which contain various algorithms that are commonly used for analysis. Other specific packages can be downloaded and installed in the package region of the architecture. All these packages are combined together into meaningful libraries and are stored in the R libraries layer. Programs also use Windows libraries to access the system storage to fetch data and to store outputs.

Fig. 1 Shows the system architecture

B. Sequence Diagram

Figure below shows a sequence diagram. It is used to depict relationship between objects and the order in which they interact. The horizontal rows in the figure represent the messages exchanged between objects in the order.

Fig 2 Shows the sequence diagram

At the first stage, the data set obtained from the source is refined to remove null values, values containing meaningless numerical values such as negative values for quantity variable. At the second stage, the dataset is restricted to have revenue of less than $30 per transaction and a subset is taken into account. This is because the retail business taken into consideration is a small-scale gifts shop in UK. At the third stage, the data is partitioned into training and testing sets. Algorithms and analytical techniques are applied to the data in the fourth stage. The result output from the model is then used for comparative study of the algorithms run on the data. The accuracy of prediction, churn values and the churn rate is then displayed to the user at the fifth stage.

C. Data Flow Diagram

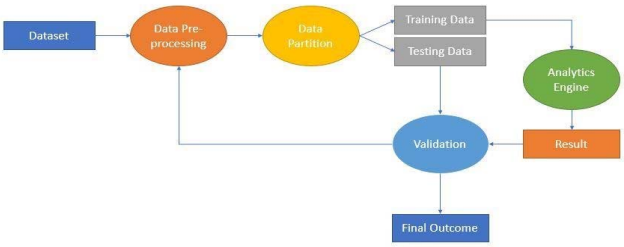

Fig 3. Shows the data flow diagram of the system

A data flow diagram, also known as DFD, is used to show a graphical representation of the data flow through the system. As described in the figure, the dataset is first pre-processed in order to get rid of null and abnormal values. This processed data is then input to the data partition module where the data is split into training and testing set. The training set is then passed to the analytics engine where a model is devised based on one of three algorithms taken into consideration. The training is basically used to train the model to predict better results. The model thus obtained is provided with test dataset as input and the results thus obtained are validated with the already existing results. If the results are satisfactory, the model is accepted. Otherwise the data is resent for pre-processing.

IV. Implementation

A. Tools

R: Reference [6] defines R as an open source programming language and a software environment that is used for statistical computations and graphical representation of the same. It was developed under GNU project. It is being widely used among data miners and statisticians who perform data analysis. It contains a vast collection of statistical and graphical techniques. The statistical techniques include linear and non-linear modelling, time-series analysis, clustering, classification and many others. The speciality of graphical tool provided with R is that, it can produce publication quality plots with little or no difficulty. Hence, the user has the full control over the functioning of the program and the visualization. R compiles and runs on all Unix platforms and also Windows and MacOS.

Tableau: As per [7], it is a software that helps user represent data through proprietary visualization tool called VizQL. It is a visual query language which converts actions in the form of drag-and-drop into data queries. This helps users to quickly analyse and share insights in their data. Tableau is first connected to data stored in files, cubes, databases or warehouses. The users then work on the Tableau UI to simultaneously query the data and view the results in the form of graphs, maps and charts.

B. Overall View of Implementation

There are primarily 5 major tasks involved in the implementation stage.

Data Preparation: The first step of predicting a churn customer is to collect data and pre-process it in the form required as an input to the algorithm. The CSV file is imported, cleaned, aggregated as per the requirements, and made into a new data frame. The retail sales dataset has a total size of 349096 tuples when compared to the original raw dataset which had around 541909 tuples. The pre-processing step involved cleaning of data by removing the NA values and also defining of one more attribute called Revenue which is the product of Quantity and Unit price. This process also involved creation of two more datasets using aggregation which are customer aggregate data and invoice aggregate data.

Training: The data is now partitioned and training models are developed using the training data. This model can be used to predict the churn values in the test data set.

Testing: The models developed using the training data set is now run over the testing dataset. RF, SVM and XGB models try to compute and predict the accuracy and other statistical variables.

Presenting the output: After the prediction is complete, the prediction set is plotted graphical to get a better visualization.

C. Algorithms

Random Forest: Random Forests are a group of decision trees built by taking a subset of attributes or tuples as input for each tree according to [8]. Then it uses a bagging and bootstrapping algorithm to ensemble the results obtained [9].

Support Vector Machine (SVM): SVMs are a type of models that come under supervised learning. Consider a set of training examples as in [10]. Let each record be denoted as belonging to one of the two classes. The SVM constructs a model that can take new examples and categorize them into one of the two categories. Hence, it is a binary linear classifier which is binary and non-probabilistic type.

Gradient Boosting: Gradient boosting is a machine learning technique for regression and classification problems [11]. It generates a model for prediction as an ensemble of prediction models that are relatively weak, typically decision trees. The model is built stage-wise just like other boosting algorithms as mentioned in [12].

D. Modules

- main.r is the main script which calls other scripts. There are totally five scripts that are called from the main script.
- preprocess.R: This script takes a raw csv file as an input and converts into a data frame. Further, this dataset is cleaned and refined by removing NAs and validating numerical quantities' ranges. Further, a new binary numerical variable called churn is added to the data points. Churn variable determines if a customer is a churn (1) or not (0).
- invagg.R: This script performs invoice based aggregations. Each invoice is processed and added to the data frame.
- custagg.R: This script performs customer based aggregations.
- training.R: This file splits data as testing and training partitions in the ratio of 0.2:0.8. It applies cross validation models on the data using Random Forest, SVM and XGB.
- testing.R: This script collects all the models together and resamples them. A dot chart is drawn explaining

the accuracy and kappa variables for each algorithm run on the training data. Further, the models modelled before are now run on testing data to predict the churn value. Accuracy of each algorithm is measured. Confusion matrix and other statistical variables are computed.

V. TESTING

The project requires specific hardware and software components for its functioning. The testing process involves unit testing, integration testing and system testing of the log analyzer. Specific setup is required for the different types of testing. Testing was done on a HP Pavillion laptop, connected to a power supply with the following specifications:

Hardware Specifications for I/O Trace and Build System:

- Intel Core i3™ CPU
- 4 GB RAM
- x64 architecture
- 500 GB HDD

Software Specifications for I/O Trace and Build System:

- Windows 10 (64-bit)
- R 3.3.3

A. Types of Testing Performed

Unit Testing: In this type of testing, a smallest segment of software that is testable in the application is isolated from the remainder of the program and is then evaluated for its behaviour. It is usually performed by the programmers themselves than the testers since it demands in-depth understanding of program's internal design.

- Local data structures: Local data that is stored in the module is tested for its proper storage in this method.
- Independent paths: This is used to ensure that every path that is independent is executing their task properly and terminating as expected as the program ends.
- Error handling paths: Error handling is a major aspect of the program and these help in reviewing if the same is handled correctly or not.

Integration Testing: Multiple test modules that have undergone the unit testing are then combined to form subsystems. These are then tested to verify correct module integrations.

Regression Testing: At times, when a code is modified for feature enhancement, the existing system may behave in an unexpected way. This test is performed to ensure that such modifications do not have any undesired impact on the existing system.

Performance Testing: The speed and efficiency of a program is evaluated using Performance testing. This test can also measure other attributes such as reliability, resource utilization and scalability.

B. Conclusion

The test cases were executed successfully and thus the project was tested and modified accordingly when the changes were required or when test cases failed.

VI. RESULTS

```
[1] "RANDOM FOREST..."
note: only 2 unique complexity parameters in default grid. Truncating the grid to 2 .

[1] "Finished training RF model in 46.75 seconds. CPU Time = 2.2 seconds."
[1] "SUPPORT VECTOR MACHINE..."
[1] "Finished training SVM model in 167.03 seconds. CPU Time = 9.47 seconds."
[1] "XTREME GRADIENT BOOSTING..."
[1] "Finished training XGB model in 141.95 seconds. CPU Time = 44.03 seconds."
[1] "Testing models over data..."
[1] "RF Accuracy:  0.652644230769231"
[1] "SVM Accuracy:  0.701923076923077"
[1] "XGB Accuracy:  0.717548076923077"
[1] "Finshed testing. Check on the right for results."
>
```

Fig. 4 Shows code execution

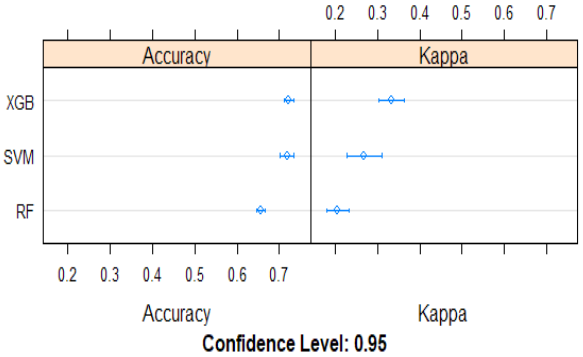

Fig. 5 Dot plot predicting accuracy and kappa values for algorithms

```
> summary(results)

Call:
summary.resamples(object = results)

Models: RF, XGB, SVM
Number of resamples: 10

Accuracy
     Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
RF  0.6317  0.6482 0.6597 0.6561  0.6667 0.6737    0
XGB 0.6976  0.7133 0.7267 0.7224  0.7333 0.7417    0
SVM 0.6856  0.7035 0.7201 0.7182  0.7273 0.7568    0

Kappa
      Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
RF  0.1385  0.1875 0.2073 0.2061  0.2368 0.2489    0
XGB 0.2788  0.2937 0.3364 0.3337  0.3754 0.3817    0
SVM 0.1888  0.2197 0.2777 0.2691  0.3110 0.3571    0
```

Fig. 6 Shows the summary of the results

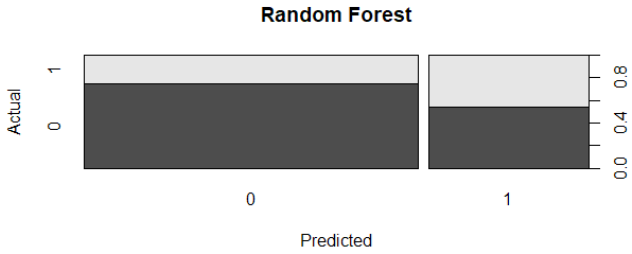

Fig. 7 shows the predictions of Random Forest algorithm

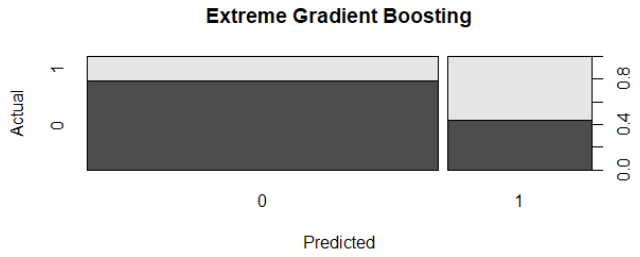

Fig 8. shows the predictions of Support vector machine

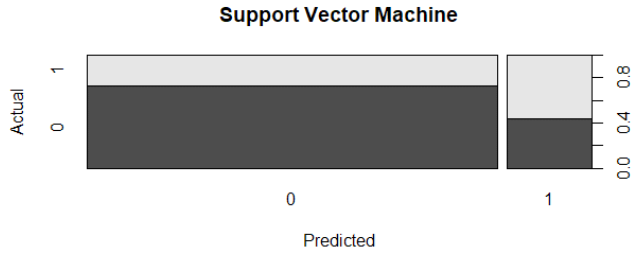

Fig. 9 shows the predictions of Extreme gradient boosting

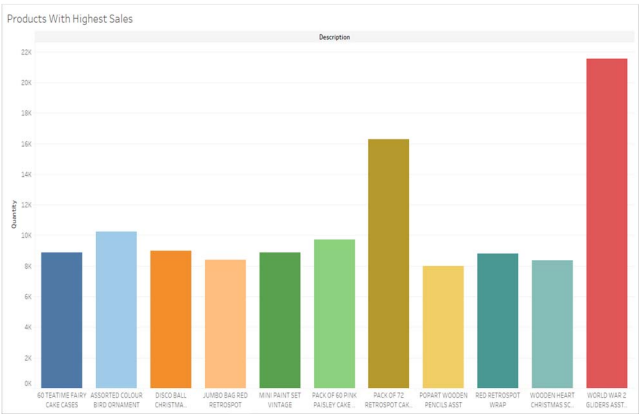

Fig. 10 Shows the products with highest sales

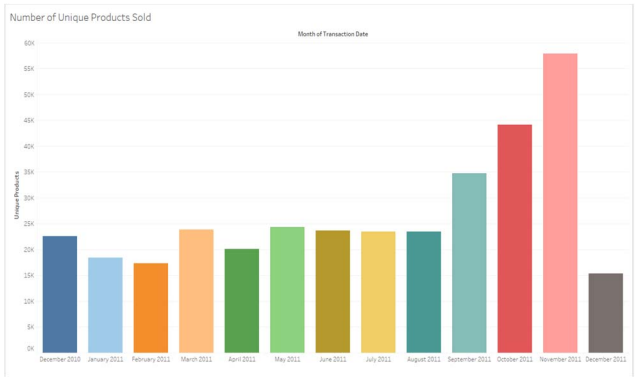

Fig. 11 Shows the total number of unique products sold in different months

TABLE I.
COMPARISON OF RF, SVM AND XGB ALGORITHMS

Parameters	Values Obtained		
	RF	SVM	XGB
Accuracy	0.6526	0.7019	0.7175
Kappa	0.2039	0.2174	0.334
Sensitivity	0.7407	0.8889	0.8131
Specificity	0.4642	0.3019	0.5132
Prevalence	0.6815	0.6815	0.6815
Detection Rate	0.5048	0.6058	0.5541
Detection Prevalence	0.6755	0.8281	0.7091
Balanced Accuracy	0.6024	0.5954	0.6631

VII. SCOPE AND FUTURE WORK

The project is currently able to predict the churn variable with an accuracy of 65%, 70% and 71% respectively for RF, SVM and XGB models for the given dataset. Therefore, this can be used for customer churn prediction for a retail business with a reasonable accuracy and time.

As already mentioned, the project is currently able to predict churn. This can further be developed to provide recommendations for a churn customer to make him a loyal customer again. For example, a customer identified to be a churn customer for a product X, can be given attractive offers and discounts to retain the customer.

VIII. CONCLUSION

This project as already stated in the objective aims to perform a comparative analysis of algorithms in predicting customer churn for a retail business. This project helped us understand different prediction algorithms in-depth. We analyzed each algorithm, studied pros and cons of each and compared them among each other. During our prediction modelling, we earned insights into which factors to take into consideration among other things.

After running all the three models, we arrived at a result that accuracy of Random Forest, Support Vector Machines and Extreme Gradient Boosting are increasingly higher. However, the time taken to finish computation also increases in the same order.

Acknowledgment

We would like to heartily thank Dept. of CSE, Ramaiah Institute of Technology for their continuous support and adequate infrastructure. We would like to express our heartfelt thanks to all the teaching and non-teaching faculty and our dear friends who helped in bringing out this paper.

References

[1] Au, W. H., Chan, K., & Yao, X, "A novel evolutionary data mining algorithm with applications to churn prediction.," IEEE Transactions on Evolutionary Computation, (2003). 7(6), 532–545.

[2] Eiben, A. E., Koudijs, A. E., & Slisser, F., "Genetic modelling of customer retention. Lecture Notes in Computer Science," (1998). 1391, 178–186.

[3] Mozer, M. C., Wolniewicz, R., & Grimes, D. B., "Churn reduction in the wireless industry," Advances in Neural Information Processing Systems, (2000). 12, 935–941.

[4] Burez, J., & Van den Poel, D., "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services," Expert Systems with Applications, (2007). 32(2), 277–288.

[5] Chen, C., Liaw, A., & Breiman L., "Using random forests to learn imbalanced data," Technical Report 666. Statistics Department of University of California at Berkeley, (2004).

[6] "RStudio", *En.wikipedia.org*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/RStudio.

[7] "Step by Step Resource Guide To Learn Tableau", Analytics Vidhya, 2017. [Online]. Available: https://www.analyticsvidhya.com/learningpaths-data-science-business-analytics-business-intelligence-bigdata/tableau-learning-path/

[8] Lemmens, A., & Croux, C., "Bagging and boosting classification trees to predict churn," DTEW Research Report, (2003). 0361.

[9] Liaw, A., & Wiener, M., "Classification and regression by random forest," The Newsletter of the R. Project, (2002). 2(3), 18–22.

[10] "Introduction to Support Vector Machines — Opencv 2.4.13.2 Documentation". *Docs.opencv.org*. [Online] Available: http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[11] "Learn Gradient Boosting Algorithm for Better Predictions (With Codes In R)". *Analytics Vidhya*. [Online] Available: https://www.analyticsvidhya.com/blog/2015/09/completeguide-boosting-methods/

[12] "Gradient Boosting". *En.wikipedia.org*. [Online] Available: https://en.wikipedia.org/wiki/Gradient_boosting