Customer Retention Project

Submitted by:

Laxmi Narayan

Acknowledgement

I would like to express my gratitude to my primary SME, Sajid Choudhary, who guided me throughout this project. I would also like to thank my friends and family who supported me and offered deep insight into the study. I wish to acknowledge the help provided by the technical and support staff in the **Data Science of Flib Robo Technologies**. I would also like to show my deep appreciation to my supervisors who helped me finalize my project.

Introduction

A good method for companies to increase revenue is to invest in pre-existing customers. Retaining customers becomes essential for consumer dependent businesses. If customers are on the verge of leaving the service, it would be prudent for the company to identify what factors influence a customer's lack of interest in the product. As the cost of retaining a customer is far lesser than getting a new one, analyzing customer churn can reveal valuable insights. The loss of customers is known as customer churn or customer attrition. Companies analyze customer churn to uncover which factors lead to a customer voluntarily switching to a rival business.

Customer churn occurs when customers or subscribers stop doing business with a company or service. Also known as customer attrition, customer churn is a critical metric because it is less expensive to retain existing customers than it is to acquire new customers – earning business from new customers means working leads all the way through the sales funnel, utilizing your marketing and sales resources throughout the process. Customer retention, on the other hand, is generally more cost-effective, as you have already earned the trust and loyalty of existing customers.

Using a machine learning approach, we can find patterns which cause customer churn and forecast it to obtain a prognosis on which factors impact customer retention. It can help make sense of relationships between data. The models can predict customers with high probability to churn based on analyzing customers personal, demographic and behavioural data to provide personalized and customer-oriented marketing campaigns to gain customer satisfaction.

Most of comparisons in the literature did not consider a study that covers the various categories of learning techniques.

The lifecycle of business – customer relationship includes four main stages: 1) *identification*; 2) *attraction*; 3) *retention*; and 4) *development*.

1) Customer identification/acquisition: This aims to identify profitable customers and the ones that are highly probable to join organization. Segmentation and clustering techniques can explore customers' personal and historical data to create segments/sub-groups of similar customers [1], [2].

2) Customer attraction: The identified customer segments / sub-groups are analyzed to identify the common features that distinguish customers within a segment. Different marketing techniques can be used to target different customer segments such targeted advertising and/or direct marketing [3].

3) Customer retention: This is the main objective of CRM as retaining existing customers is at least 5 to 20 times more cost effective than acquiring new ones depending on business domains [4], [5]. Customer retention includes all actions taken by organization to guarantee customer loyalty and reduce customer churn. Customer churn refers to customers moving to a competitive organization or service provider. Churn can be for better quality of service, offers and/or benefits. Churn rate is an important indicator that all organizations aim to minimize. For this sake, churn prediction is an integral part of proactive customer retention plan [6]. Churn prediction includes using data mining and predictive analytical models in predicting the customers with high likelihood to churn/defect. These models analyze personal and behavioral customer data for tailored and customer-centric retention marketing campaigns [7].

4) Customer development: The main objective of this phase is to increase the amount of customer transactions for more profitability. For this sake, market basket analysis, customer lifetime value, up, and cross selling techniques are used. Market basket analysis tries to analyze customers' behavior patterns to maximize the intensity of transactions [8], [9]. Analyzing customer lifetime value (CLTV) can help identifying the total net income expected from customer [10]-[12]. Up and/or Cross selling include activities that increase the transactions of the associated services/products [13], [14].

Most of comparisons in the literature did not consider a study that covers the various categories of learning techniques. The bulk of the models applied for churn prediction fall into one of the following categories:

1) Regression analysis, 2) Decision tree–based, 3) Support Vector Machine, 4) Bayesian algorithm, 5) Instance – based learning, 6) Ensemble learning, 7) Artificial neural network, and 8) Linear Discriminant Analysis.

This study presents a comparative study of the most used algorithms for predicting customer churn. The comparison is held between algorithms from different categories. The main goal is to analyze and benchmark the performance of the models in the literature. The selected models are:

1) Regression analysis: logistic regression.
2) Decision tree–CART.
3) Bayes algorithm: Naïve Bayesian.
4) Support Vector Machine
5) Instance – based learning: k-nearest Neighbor.
6) Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest.
7) Artificial neural network: Multi-layer Perceptron.
8) Linear Discriminant Analysis.

A. Contribution

The key contribution of this paper is the analysis of most common learning techniques in the state of the arts and the evaluation of their accuracy.

Machine-Learning For Churn Prediction

Machine-learning techniques have been widely used for evaluating the probability of customer to churn [25]. Based on a survey of the literature in churn prediction, the techniques used in the bulk of literatures fall into one of the following categories 1) Regression analysis; 2) Tree – based; 3) Support Vector Machine; 4) Bayesian algorithm; 5) Ensemble learning; 6) Sample – based learning; 7) Artificial neural network; and 8) Linear Discriminant Analysis. A brief introduction of the chosen algorithms is presented in this section.

1) Regression analysis: Regression analysis techniques aim mainly to investigate and estimate the relationships among a set of features. Regression includes many models for analyzing the relation between one target/response variable and a set of independent variables. Logistic Regression (LR) is the appropriate *regression analysis* model to use when the dependent variable is binary. LR is a predictive

analysis used to explain the relationship between a dependent binary variable and a set of independent variables. For customer churn, LR has been widely used to evaluate the churn probability as a

2) Decision Tree: Decision Tree (DT) is a model that generates a tree-like structure that represents set of decisions. DT returns the probability scores of class membership. DT is composed of: a) **internal Nodes:** each node refers to a single variable/feature and represents a test point at feature level; b) **branches,** which represent the outcome of the test and are represented by lines that finally lead to c) **leaf Nodes** which represent the class labels. That is how decision rules are established and used to classify new instances. DT is a flexible model that supports both categorical and continuous data. Due to their flexibility they gained popularity and became one of the most commonly

3) Support Vector Machine: Support Vector Machine (SVM) is a supervised learning technique that performs data analysis in order to identify patterns. Given a set of labeled training data, SVM represents observations as points in a high-dimensional space and tries to identify the best separating hyperplanes between instances of different classes. New instances are represented in the same space and are classified to a specific class based on their proximity to the separating gap. For churn prediction, SVM techniques have been widely investigated and evaluated to be of high predictive performance [37]-[41].

4) Bayes Algorithm: Bayes algorithm estimates the probability that an event will happen based on previous knowledge of variables associated with it. Naïve Bayesian (NB) is a classification technique that is based on Bayes' theorem. It adopts the idea of complete variables independence, as the presence/absence of one feature is unrelated to the presence/absence of any other feature. It considers that all variables independently contribute to the probability that the instance belongs to a certain class. NB is a supervised learning technique that bases its predictions for new instances based on the analysis of their ancestors. NB model usually outputs a probability score and class membership. For churn problem, NB predicts the probability that a customer will stay with his service provider or switch to another one [42]-[46].

5) Instance – *based learning*: Also known as ***memory-based learning***, new instances are labeled based on previous instances stored in memory. The most widely used instance based learning techniques for classification is K-nearest neighbor (KNN). KNN does not try to construct an internal model and computations are not performed until the classification time. KNN only stores instances of the training data in the features space and the class of an instance is determined based on the majority votes from its neighbors. Instance is labeled with the class most common among its neighbors. KNN determine neighbors based on distance using Euclidian, Manhattan or Murkowski distance measures for continuous variables and hamming for categorical variables. Calculated distances are used to identify a set of training

instances (k) that are the closest to the new point, and assign label from these. Despite its simplicity, KNN have been applied to various types of applications. For churn, KNN is used to analyze if a customer churns or not based on the proximity of his features to the customers in each classes [17], [51].

6) Ensemble – *based Learning*: Ensemble based learning techniques produce their predictions based on a combination of the outputs of multiple classifiers. Ensemble learners include bagging methods (i.e. Random Forest) and boosting methods (i.e. Ada Boost, stochastic gradient boosting).

***a)* Random Forest**

Random forests (RF) are an ensemble learning technique that can support classification and regression. It extends the basic idea of single classification tree by growing many classification trees in the training phase. To classify an instance, each tree in the forest generates its response (vote for a class), the model choses the class that has receive the most votes over all the trees in the forest. One major advantage of RF over traditional decision trees is the protection against overfitting which makes the model able to deliver a high performance [47]-[50].

***b)* Boosting – based techniques (Ada Boost and Stochastic Gradient Boosting)**

Both AdaBoost (Adaptive Boost) and Stochastic Gradient Boosting algorithms are ensemble based algorithms that are based on the idea of boosting. They try to convert a set of weak learners into a stronger learner. The idea is that having a weak algorithm will perform better than random guessing. Thus, Weak learner is any algorithm that can perform at least a little better than random solutions. The two algorithms differ in the iterative process during which weak learners are created. Adaboost filters observations, by giving more *weight* to problematic ones or the ones that the weak learner couldn't handle and decrease the correctly predicted ones. The main focus is to develop new weak learns to handle those misclassified observations. After training, weak learners are added to the stronger learner based on their alpha weight (accuracy**)**, the higher alpha weight, the more it contributes to the final learner. The weak learners in AdaBoost are decision trees with a single split and the label assigned to an instance is based on the combination of the output of all weak learners weighted by their accuracy [56].

7) Artificial neural network: Artificial Neural Networks (ANNs) are machine-learning techniques that are inspired by the biological neural network in human brain. ANNs are adaptive, can learn by example, and are fault tolerant. An ANN is composed of a set of connected nodes (neurons) organized in layers. The input layer communicates with one or more hidden layers, which in turn communicates with the output layer. Layers are connected by weighted links. Those links carry signals between neurons usually in the form of a real number. The output of each neuron is a function of the weighted sum of all its inputs.

The weights on connection are adjusted during the learning phase to represent the strengths of connections between nodes. ANN can address complex problems, such as the churn prediction problem. Multilayer perceptron (MLP) is an ANN that consists of at least three layers. Neurons in each layer use supervised learning techniques [52], [53]. In the case of customer churn problem, MLP has proven better performance over LR [21], [27], [28], [54], [55].

8) Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) is a mathematical classification technique that searches for a combination of predictors that can differentiate two targets. LDA is related to regression analysis. They both attempt to express the relationship between one dependent variable and a set of independent variables. However, unlike regression analysis, LDA use continuous independent variables and a categorical dependent variable (target). The output label for an instance is estimated by the probability that inputs belong to each class and the instance is assigned the class with the highest probability. Probability in this model is calculated based on Bayes Theorem. LDA can be used for dimensionality reduction by determining the set of features that are the most informative. LDA has been used in for different classification tasks including customer churn [63]-[65].

METHODOLOGY

1) Data: The used dataset for the experiments of this study is a database of customer data of a telecommunication company. There are two sheets (one is detailed) and second is encoded in the excel file. You may use any of them by extracting in separate excel sheet. The number of column(s) is more than 47..

2) Data preprocessing: Preprocessing includes three steps: a) data transformation, b) data cleaning and c) feature selection.

a) Data Transformation

Data transformation is the process of converting data from one format to another. Data transformation can increase the efficiency of analytic and business processes and enable better data-driven decision-making.

b) Data cleaning

This stage includes missing data handling/imputation: Some of the selected algorithms cannot handle missing data such as SVM. That's why missing value can be replaced by mean, median or zero. However, missing data replacement by statistically computed value (imputation) is a better option. The used dataset included missing values in some the numerical variables (Day Charge, Eve Mins, Intl Calls, Intl Charge and Night Charge) and two categorical variables (VMail Plan, Int'l Plan). Numerical data were replaced using random forest imputation technique [69]. And binary values were imputed using the techniques in [70]

c) Feature selection

Before model training, feature selection is one of the most important factors that can affect the performance of models. In this study, the importance of the used variables was measured to identify and rank explanatory variables influence on the target/response. This allows dimensionality reduction by removing variables/predictors with low influence on the target. Random forest technique can be used for feature selection using mean decrease accuracy. Mean decrease measures the impact of each feature on model accuracy. The model permutes values of each feature and evaluates model accuracy change. Only features having higher impact on accuracy are considered important [71]

3) Simulation Setup: For this study, the selected models are used to generate predictions using the dataset containing 3333 samples with 13 predictors and one response variable. 10-fold cross validations were used for models training and testing. Training and testing datasets are randomly chosen with cross validation 70% for training and 30% for testing. Each module requires initial parameters that are set as follows:

a) Decision Tree (CART)

One parameter is used for decision tree, CP which is a complexity parameter used to control the optimal tree size. Accuracy is used to choose the optimal model. The final (cp) value used for the model was: 0.07867495

b) Support Vector Machine

In order to train SVM, two main parameters are required: C and Sigma. The C parameter affects the prediction. It indicates the cost of penalty. Large value For C means high accuracy in training and low accuracy in testing. While small value for C indicates unsatisfactory accuracy. While sigma parameters has a more influence than C on classifications, as it affects hyperplane partitioning. A too large value of sigma leads to over-fitting, while small values lead to under-fitting [73]. Cross-validation was performed to select and tune performance parameters. The values that gave the highest accuracy were sigma = 0.06295758 and C = 1 as shown in Table IV.

c) K-nearest Neighbor

In KNN, one parameter needs to be tuned. K is the number of instances/neighbors that are considered for labeling an instance to a certain class. Cross validations were performed using different k values. Results shown in Table V shows that the highest accuracy is obtained using k=7.

***d)* AdaBoost**

For Ada boost mode, *nIter* - represents the number of weak learners to be used. Grid search was used to determine the best accuracy. Results show that highest accuracy is at nIter=100.

***e)* Random Forest**

A forest of 500 decision trees has been built using the Random Forest algorithm. Error rate results indicate that after 100 trees, there is no significant error reduction. Another parameter is *mtry that* indicates number of predictors sampled for spliting at each node. Results in Table VII show that the optimal performance is at mtry = 10.

This study presents a comparative study of the most used algorithms for predicting customer churn. The comparison is held between algorithms from different categories. The main goal is to analyze and benchmark the performance of the models in the literature

Online retail customers values

This category of e-retail customer would be motivated to shop on an e-vendor with widely selection of products, informative, convenient, and fast websites (Sorce et al., 2005). Both categories of shopping values have been presented in Fig. 1

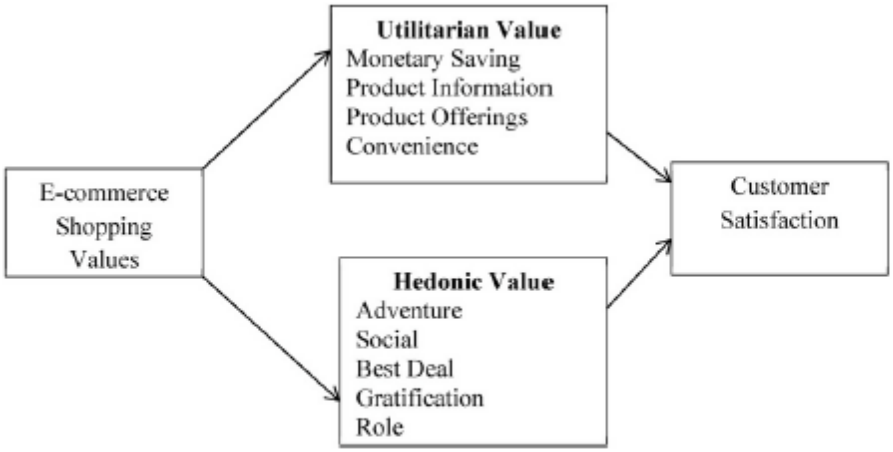

Fig. 1. E-commerce shopping values.

Proposed Customer Retention Model

This research study adopts the theory that an online customer's repeat purchase intention (Retention) is influenced by utilitarian and hedonic values, which are further derived from net benefits (Childers et al., 2001; Sorce et al., 2005). This study presents the second order composite latent variables in the customer retention model.

The research models utilised in the research attempts to investigate why benefits are considered as the components of values; and to establish the link between goals and value. The customer retention model takes its root from the Means-End Chain theory (MEC) pay particular attentions to "consumers' perceptive state after product or service consumption, it does not involve the risk concept.

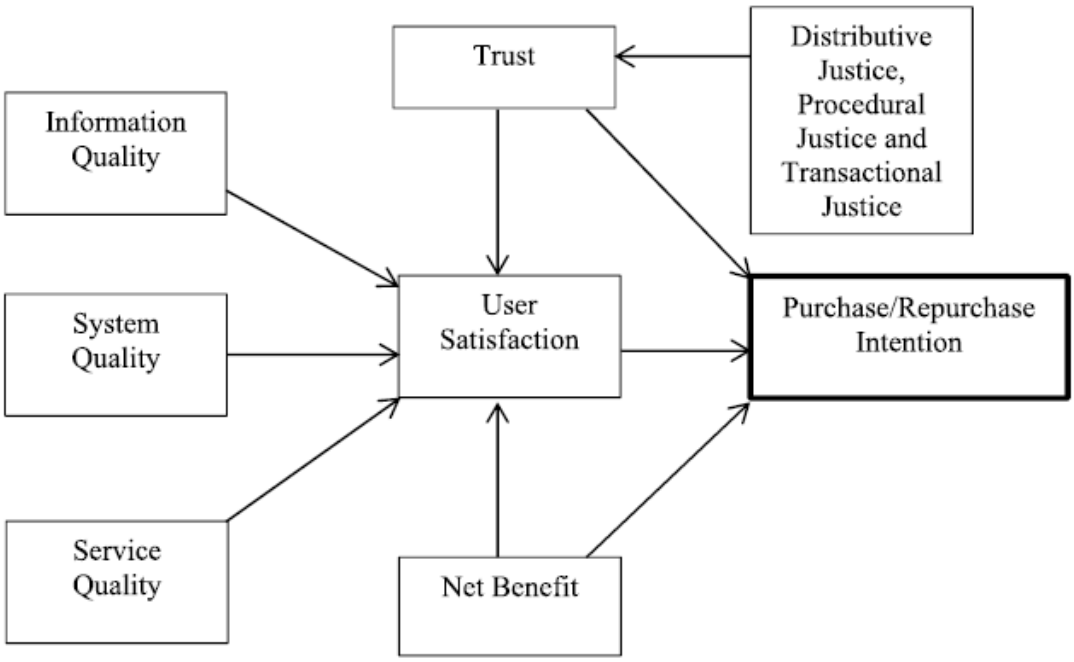

Fig2: Proposed customer activation model

Data analysis

Aim of the data analysis was to test and validate the proposed conceptual model and to identify the casual relationships that exist between variables. The researches adopted several techniques to test, validated, check the fitness of the model as well as to test the hypotheses. The study used several goodness-of-fit indices to test if the model can be accepted; the technique used includes; Chi-Square Goodness of Fit Test, Normed Fit Index (NFI), Comparative Fit Index (CFI), The Root Mean Square Error of Approximation (RMSEA), and Tucker-Lewis index (TLI).

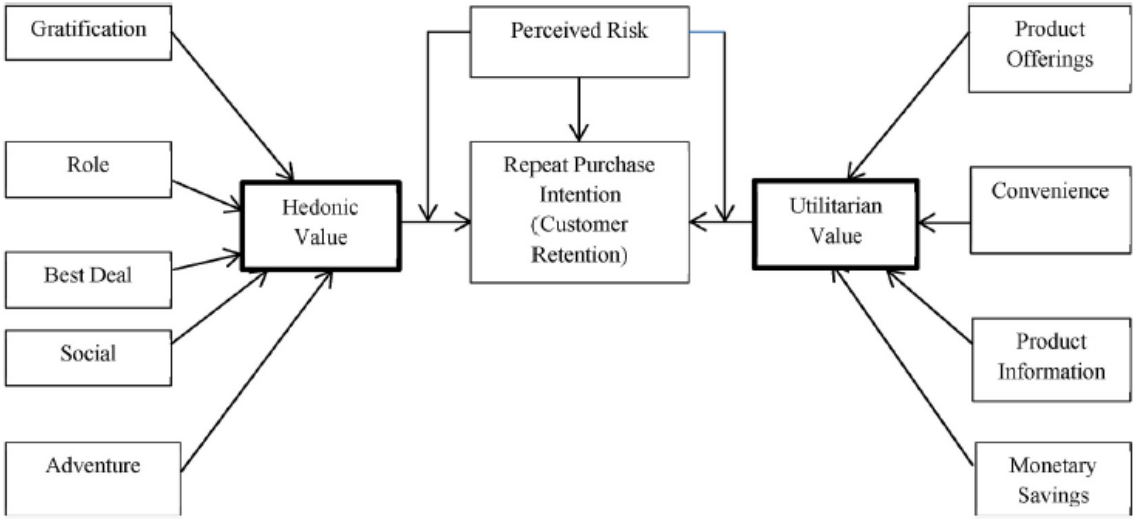

Fig3:Proposed customer retention model

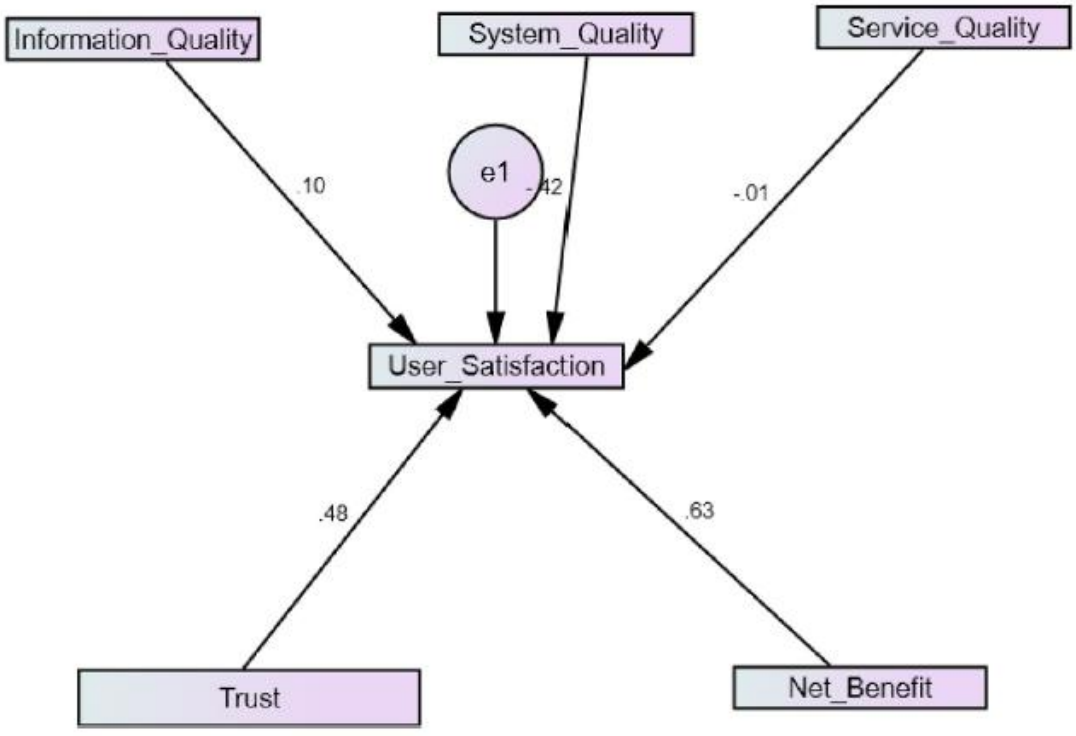

Fig4: Structural equation model of customer activation model.

Literature Survey

1) The Analysis of Customer Churns in E-commerce based on Decision Tree [1]

Published in: 2015 International Conference on Computer Science and Applications (CSA)

Summary: The authors of this paper used the Decision Trees method to analyze an e-commerce's customer churn. They have approached this problem by using classification to predict customer churn. Using this method, they were able to find important features that attribute to attrition. They considered this method suitable for classification because of its straightforward approach. Their model has 88% prediction accuracy. They were able to discover that the discount revenue rate being less than 42.3% of average, classified the customer to be in sleep status. This provided us with an insight into how researchers approached customer churn.

Strengths and weaknesses: Decision tree model presentation and results are understandable by the stakeholders and is an intuitive approach to predicting customer churn. Their approach focuses on the the effect of using decision trees on their e-commerce dataset.

Relation to approach: We do not plan to limit ourselves to one model. Our approach will use decision trees to contrast its result with other machine learning models. Our focus is on the features present and developing relevant features to augment the dataset present and improve the results.

2) Telecommunication Subscribers' Churn Prediction Model Using Machine Learning [2]
Published in: Eighth International Conference on Digital Information Management (ICDIM 2013)

Summary: The researchers used Linear Regression, Logistic Regression, Artificial Neural Networks, K-Means Clustering and Decision Trees variations (CHAID, Exhaustive CHAID, CART, QUEST) to classify their 106,000-sample dataset from Customer DNA's website. Active and churn customers were classified using these models. In their literature survey covered paper which used Neural Networks and Regression models to predict customer churn. They faced class imbalance in their dataset, i.e., there were more active users than churn users. They used various re-sampling methods to handle this imbalance. Exhaustive CHAID model delivered the best results to predict customer churn.

Strengths and Weaknesses: Customer churn is exclusively a classification problem in this paper. The imbalance present in their dataset provided insight into handling datasets with skewed classes.

Relation to approach: This paper focused on classification models and their results on an e-commerce dataset. We plan to perform regression along with classification.

3) **Machine-Learning Techniques for Customer Retention: A Comparative Study [3]**
Published in: International Journal of Advanced Computer Science and Applications (IJACSA) 2018

Summary: The author's aim was to find a benchmark for churn classification using multiple machine learning model approaches. Comparing a host of models (Regression analysis: logistic regression, Decision tree–CART, Bayes algorithm: Naïve Bayesian, Support Vector Machine, Instance-based learning: k-nearest Neighbor, Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest, Artificial neural network: Multi-layer Perceptron, Linear Discriminant Analysis) on a telecommunication dataset with 3333 records. Evaluating each model's performance using model metrics, they have provided an in-depth analysis of how each model performs classification to find the customer churn.

Strengths and weaknesses: This comparative study provides us information on how each model performs on their low sample data.

Relation to approach: SVM gave the best results and could be considered as a contrast model in our approach to finding the best results for customer churn. They used K-fold cross-validation to present the results for each model. Our approach plans to focus on improving the feature set as opposed to a comparative study which is presented in this paper.

4) **Behavioural Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty [4]**
Published in: 2015 IEEE International Congress on Big Data

Sumarry: The authors of this paper developed a "churn score" which they assigned to each customer to predict the early signs of customer churn to indicate the likelihood of customer churn. They used feature

engineering to find the most relevant features, assign a probability(churn score) to each customer and subsequently input the features into different machine learning models.

Strengths and Weaknesses: This paper is closest to what we want to achieve from our project. Their feature engineering and feature selection presented an approach to a feature engineering focused analysis. They have a vastly different dataset with anonymous feature names, which they had to extract and build new features based on a supervised learning approach to finding a retention score based on probability.

Relation to approach: The dataset used is vast with over a billion samples and multiple features. They focus on subscription data of prepaid customers in a telecom dataset. It requires feature engineering and selection to find the best features to input into machine learning models. Our approach aims to use survival analysis to find the retention score of each customer as opposed to the inactive day's parameter used in this paper.

Results and discussion

Accuracy is used to evaluate the model performance. Accuracy indicates the ability to differentiate the credible and non-credible cases correctly. It's the proportion of true positive (TP) and true negative (TN) in all evaluated news:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

```
Accuracy Score of Logistic Regression model is 100.0
Accuracy Score of Decision Tree Classifier model is 100.0
Accuracy Score of K-Nearest Neighbour Classifier model is 100.0
Accuracy Score of Support Vector Classifier model is 100.0
Accuracy Score of Random Forest model is 100.0
Accuracy Score of ADA Boost model is 61.72839506172839
```

Future Prediction

```
Accuracy Score of RFC model is 100.0
Confusion matrix for RFC Model is
[[26  0  0  0  0  0  0  0]
 [ 0 22  0  0  0  0  0  0]
 [ 0  0  4  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  0]
 [ 0  0  0  0  5  0  0  0]
 [ 0  0  0  0  0  7  0  0]
 [ 0  0  0  0  0  0 11  0]
 [ 0  0  0  0  0  0  0  2]]
Classification Report of the RFC Model is
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        26
           1       1.00      1.00      1.00        22
           2       1.00      1.00      1.00         4
           3       1.00      1.00      1.00         4
           4       1.00      1.00      1.00         5
           5       1.00      1.00      1.00         7
           6       1.00      1.00      1.00        11
           7       1.00      1.00      1.00         2

    accuracy                           1.00        81
   macro avg       1.00      1.00      1.00        81
weighted avg       1.00      1.00      1.00        81
```

Conclusion

The results of this study suggest following outputs which might be useful for E-commerce websites to extend their business

1. This study tries to present a benchmark for the most widely used state of the arts for churn classification. The accuracy of the selected models was evaluated on a public dataset of Customer Retention. Based on the findings of this study, ensemble – based learning techniques are recommended as both Random forest and Ad boost models gave the best accuracy.

2. However, the study can be extended by including hybrid models and deep learning models. Other performance metrics can be used for performance evaluation. Timing measures of the models can also be a major indicator for performance. Models can also evaluate against different datasets from different domains.

References

[1] Guha, Sudipto, and Nina Mishra. "Clustering data streams:- In Data Stream Management". Springer Berlin Heidelberg, 2016.

[2] Brito, Pedro Quelhas, Carlos Soares, Sérgio Almeida, Ana Monte, and Michel Byvoet. "Customer segmentation in a large database of an online customized fashion business." Robotics and Computer-Integrated Manufacturing , Elsevier ,2015.

[3] Abolfazl Kazemi, Mohammad Esmaeil Babaei, "Modelling Customer Attraction Prediction in Customer Relation Management using Decision Tree: A Data Mining Approach", journal of Optimization in Industrial Engineering , 2011.

[4] The Chartered Institute of Marketing, Cost of customer acquisition versus customer retention (2010).

[5] Colin Shaw, CEO, Beyond Philosophy, 15 Statistics That Should Change The Business World – But Haven't, Featured in: Customer Experience, June 4, 2013.

[6] Ramakrishna Vadakattu ; Bibek Panda ; Swarnim Narayan ; Harshal Godhia " Enterprise subscription churn prediction" ,IEEE International Conference on Big Data (Big Data), 2015.

[7] Miguel A.P.M. Lejeune "Measuring the impact of data mining on churn management", Internet Research , Vol. 11 Issue: 5, pp.375-387,

[8] Jain S., Sharma N.K., Gupta S., Doohan N. (2018) Business .Strategy Prediction System for Market Basket Analysis. In: Kapur P., Kumar U., Verma A. (eds) Quality, IT and Business Operations. Springer Proceedings in Business and Economics. Springer, Singapore. 2017. DOI https://doi.org/10.1007/978-981-10-5577-5_8

[9] Andrew h. Karp, using logistic regression to predict customer retention, 1998.

[10] Cheng, C.-H. and Chen, Y.-S. (2009), "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications, Vol. 36 No. 3, pp. 4176-4184.

[11] Huang, S.C., Chang, E.C. and Wu, H.H. (2009), "A case study of applying data mining techniques in an outfitter's customer value analysis", Expert System Application, Vol. 36 No. 3, pp. 5909-5915

[12] Nishant Saxena, ESCORT (Enterprise Services Cross-sell Optimization Using Rigorous Tests of Association), Advances in Economics and Business 5(5): 239-245, 2017

[13] Anita Prinzie , DirkVan den Poel, "Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models". European Journal of Operational Research, 170, 710–734. 2006.

[14] Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.2 , February 2011

[34] Luo Bin ; Shao Peiji ; Liu Juan, “Customer churn prediction based on the decision tree in personal handyphone system service”. International Conference on Service Systems and Service Management, 2007.