Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:05:2022
DOI 10.17605/OSF.IO/ZRX7H

RESEARCH ON CHURN PREDICTION IN MOBILE COMMERCE USING SUPERVISED MODEL.

Dr. CHITRA KIRAN. N
Professor, Department of Electronics and Communication Engineering, Alliance University, Bangalore INDIA

WESELY SUSHANTH VAILSHERY
Faculty of Electrical Engineering and Information Technology, Technische Universitat Chemnitz

SANDEEP A PATIL
End to End Senior Solution Professional, British Telecom UK

ABSTRACT

Churn prediction is one of the most difficult Big Data use cases. It is the most important indicator for a robust and expanding company, regardless of size or sales channel. Consumer churn, defined as a consumer leaving an established relationship with a company, is an important topic that has been extensively researched for both academic and commercial purposes. When a company's clients discontinue doing business with it, this is referred to as churn. In online commerce, a customer is considered churned when his or her transactions are outdated for more than a certain period. When a customer churns, the company suffers a loss that includes not just the lost revenue from the lost customer, but also the expenditures of further marketing to acquire new customers. The key goal of every online business is to reduce client churn. Customer attrition is detrimental to a company's bottom line. As a result, accurate customer churn prediction is critical for organizations seeking to improve client retention and corporate profits. However, there are challenges with assessing client attrition using standard methodologies in the case of mobile commerce. In this article, a Machine Learning (ML) model for predicting churn in mobile commerce is being built. By using the customer dataset, three techniques such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Naive Bayes (NB) are used to estimate churn. These three algorithms are compared using accuracy metrics in a study. When compared to other methodologies, LDA had a better churn prediction accuracy of 92.53%.

Keywords— Commerce, Mobile, Missing Values, Feature Selection, Accuracy

I. INTRODUCTION

Cell phones and tablets can be used to undertake business activities, such as purchasing and selling things, online banking, and paying bills, utilizing mobile commerce, also known as m-commerce or m-commerce. Online shopping is used in this study. When compared to typical commercial customer management, the mobile-commerce customer retention percentage is as high as 80% [1]. While studies show that recruiting new consumers is more expensive, retaining existing customers save 4 to 6 times as much money. If client retention rates rise by just 5%, operators will see an increase in profits of 85%. The basis for keeping future clients will be provided by precisely identifying consumers who churn and strengthening our ability to identify customers who don't churn. Retaining customer loyalty is a common issue in the business sector. According to one study [2] acquiring new clients is less profitable than retaining existing ones. Customer Relationship Management focuses on loyalty or churn prediction. The turnover rate is substantially lower in industries where customers are contractually obligated than in others. People in the digital age expect to buy quickly, easily, and affordably. They like to shop online because they can find a large variety of items at low prices and have them delivered directly to their house. A marketing campaign encourages people to shop online. A growing number of firms are selling their wares online. Keeping clients satisfied takes effort since they expect the best and most cheap things. Customers may quite classic to acquire a better product, resulting in a decline in sales. When this pattern of behaviour persists for an extended period, this is referred to as customer churn [3]. Customer churn prediction is a tough and time-consuming process that seeks to identify customers who are ready to discontinue using a company's product or service. Making models that can detect early churn signals and pinpoint customers who are about to leave or remain is a major objective for decision-makers and ML experts. As a result, to keep customers, it is critical to building an effective churn prediction model that reduces the risk of customer turnover. Customers' churn models have been shown by researchers to raise revenue and improve a company's market reputation. Reducing churn and retaining current customers is the most cost-effective marketing technique for increasing shareholder value. Because of the wealth of data that organizations have about their customers, the ML community may now build predictive modelling algorithms to manage the prediction of client turnover.

Several studies have been conducted in recent years to forecast customer attrition in various places. Various data mining approaches were used, resulting in a diversity of results. The next section provides a comprehensive review of previous churn prediction research. The study [5] provides customer turnover models that are tested using conventional criteria. According to the findings, ML technologies improved the proposed churn model. Random Forest and J48 both achieved a superior F-measure result of 88 percent. After identifying the key churn reasons in the dataset, we performed cluster profiling based on each individual's propensity for churning. Finally, a collection of guidelines for CEOs of telecommunications companies on how to retain clients. The paper [6] presents an e-commerce customer churn prediction model based on enhanced SMOTE and AdaBoost to improve non-churn customer identification and forecast accuracy for churn consumers. The imbalance problem is addressed using oversampling and under-sampling methodologies, and the AdaBoost algorithm is utilized to forecast the turnover rate. Finally, actual evidence from a B2C E-commerce platform shows that our model is more efficient and accurate than more mature customer churn prediction algorithms. The research [7] investigates how ML may be used to predict and analyse bank client attrition. In this data investigation, the EDA (Exploratory Data Analysis) technique is applied for the first time. Because the data is imbalanced and contains missing values, pre-processing is required. Following the selection of baseline features, a model is developed using a variety of ML approaches such as LR, DT, KNN, and RF. This article compares baseline and all-feature model comparisons. The recall, precision, AUC ROC, and accuracy are all shown. Averaging and max voting is final assembly methods used to improve model performance while sacrificing accuracy. Random Forest is the most effective model when compared to the others.

According to the author [8], client churn in the banking industry can be predicted using an LSTM model and data pre-processed using the SMOTE approach. In the financial sector, models such as the Mixed Ensemble and Hybrid models have been employed to predict customer loyalty. To accurately anticipate client attrition, the data is pre-processed using the SMOTE technique and an LSTM model. Firms are better able to identify clients who are more likely to abandon their purchases when they use this paradigm. After conducting an evaluation, it was discovered that the proposed systems for churn prediction operate with an accuracy of 88 percent, which is much higher than the accuracy of the system without the SMOTE technique. According to the article [9], an SVM-based prediction model was built as a result of the easily correlated and multi-index of suggestive elements found in the turnover data from the chain retail industry. As a result, principal component analysis (PCA) can be used to minimize the number of dimensions and eliminate redundant information, producing a more compact and acceptable collection of samples for SVM analysis. To analyse 31-dimensional feature vectors of customer turnover data in this paper, PCA was first modified before being applied to real chain retail data sets and verified to demonstrate that this PCA and SVM model outperforms alternative approaches based solely on SVM in terms of accuracy and precision. The article [10] is primarily concerned with data from the most widely used online food ordering service. Customers' churn can be identified and prevented by conducting data analysis, which can then be used to retain existing customers as well. Customers' churn prediction is the primary purpose of this essay, which makes use of information from web properties as well as information from users' activities. Ultimately, we want to know whether a user will continue to use our service in the future. In a series of trials, the findings of several data mining methodologies were compared to one another. The results reveal that, when compared to other approaches, Gradient Boosted Trees are more accurate, with an accuracy of 86.90 percent. However, while the publication [11] is primarily concerned with the prediction of mobile game churn, the presented method can be extended to a wide range of other problems. Generalizations of the suggested model can be made to address any churn prediction job where the underlying data can be described similarly, such as customer disengagement prediction in membership businesses and interest group subscription prediction in social networks The properties of objects and their relationships at various timestamps are the only information provided to the model as inputs.

The outline for this document is provided below. It first assesses recent research addressing customer churn prediction, methodologies used in this area, and data selection. Second, it offers data collection, quantitative analytical approaches for pre-processing, critical feature selection, and churn prediction modelling. Thirdly, addresses the output of the ML model as well as performance analysis. Finally, wrap up the study by emphasizing the possibility for future research as well as the conclusions.

II. RESEARCH METHODOLOGY

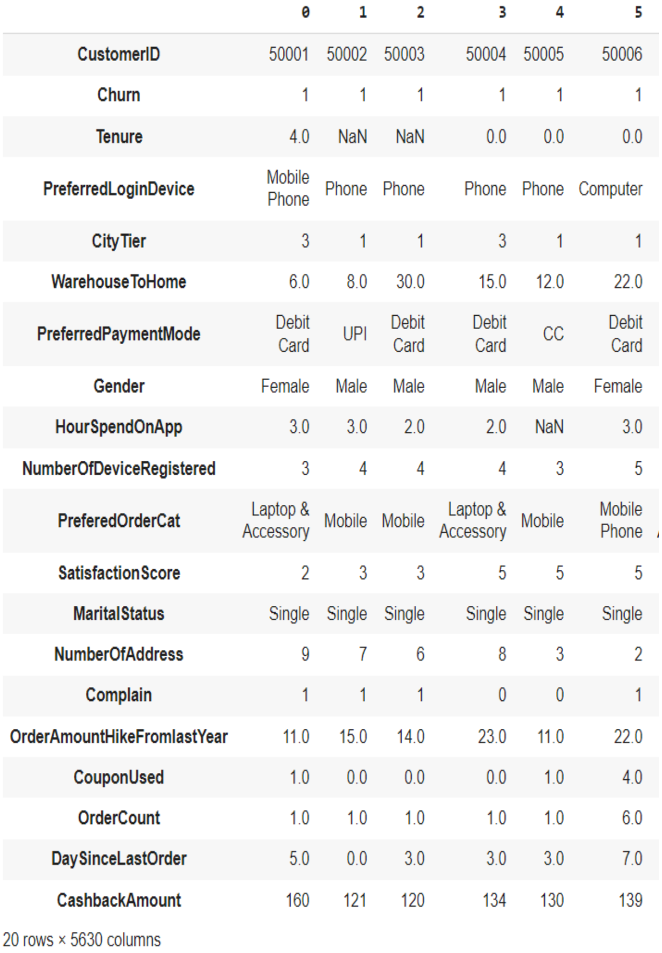
In this section, over the methodologies used to estimate client churn using mobile commerce data. Table I depicts the investigation's progression. The first column denotes the stage level, the second describes the process, and the third shows the technique used in each stage. Data collection, eradication of null values, selection of essential features, construction of an ML model, and evaluation are the steps in the process flow.

TABLE I. RESEARCH FLOW AND ITS TECHNIQUES

Stages	Process	Technique
1	Data Collection	Kaggle
2	Eliminate null values	Statistics
3	Feature selection	Heatmap
4	Model	SVM, LDA, NB
5	Evaluate	Accuracy

A. Data

The data set belongs to m-Commerce which is collected from Kaggle. The collected data contains 19 features and 1 target variable. The target variable contains 0’s and 1’s. The 0 represents non-churn as well as 1 is used for churn. The data contains 5630 instances. From 5630, 4504 is taken for train and 1126 data are used for the test. The data of 6 samples are shown in figure 1. The 1st column in the figure represents the features and targets. The first row gives customer ID, 2nd shows whether the customer churns or not. And all the rows below this show the features.



	0	1	2	3	4	5
CustomerID	50001	50002	50003	50004	50005	50006
Churn	1	1	1	1	1	1
Tenure	4.0	NaN	NaN	0.0	0.0	0.0
PreferredLoginDevice	Mobile Phone	Phone	Phone	Phone	Phone	Computer
CityTier	3	1	1	3	1	1
WarehouseToHome	6.0	8.0	30.0	15.0	12.0	22.0
PreferredPaymentMode	Debit Card	UPI	Debit Card	Debit Card	CC	Debit Card
Gender	Female	Male	Male	Male	Male	Female
HourSpendOnApp	3.0	3.0	2.0	2.0	NaN	3.0
NumberOfDeviceRegistered	3	4	4	4	3	5
PreferedOrderCat	Laptop & Accessory	Mobile	Mobile	Laptop & Accessory	Mobile	Mobile Phone
SatisfactionScore	2	3	3	5	5	5
MaritalStatus	Single	Single	Single	Single	Single	Single
NumberOfAddress	9	7	6	8	3	2
Complain	1	1	1	0	0	1
OrderAmountHikeFromlastYear	11.0	15.0	14.0	23.0	11.0	22.0
CouponUsed	1.0	0.0	0.0	0.0	1.0	4.0
OrderCount	1.0	1.0	1.0	1.0	1.0	6.0
DaySinceLastOrder	5.0	0.0	3.0	3.0	3.0	7.0
CashbackAmount	160	121	120	134	130	139

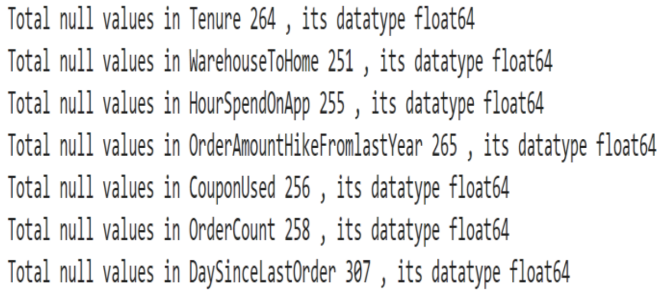
20 rows × 5630 columns

Fig. 1. Data information

B. Data Pre-processing

- Before data can be used, it must be pre-processed. The purpose of pre-processing is to convert raw data into a clean dataset. There are several key processes in data pre-processing, including data cleaning, data transformation, and feature selection [12]. Data cleaning and transformation are required to make the data more useable in the building of a model. Because ML algorithms learn from data, pre-processing data is a required step before using any algorithm. The learning outcome for problem-solving is heavily reliant on the proper data required to solve a specific challenge - known as features. Because of the relevance of these features in learning and comprehension, ML is also referred to as feature engineering. Outliers and missing numbers are typical when collecting data [13]. When there are missing values, the amount of data that can be investigated is reduced, which impacts the statistical power and, eventually, the dependability of the study. Outliers in the estimating process can cause overestimation or underestimating of values.
- It is critical to understand that "missing data" (or "missing values") refers to variables in an observation that are not included in the dataset. Many researchers encounter the problem of missing or incomplete data, which has

the potential to drastically affect their conclusions. Many scientists have made conclusions based on the assumption of extensive data collecting, which is no longer the reality. Missing data can lead to a range of problems. Missing data, as a first step, reduces statistical power, which is the likelihood that a test would reject the null hypothesis when it is false. As a second consequence, missing data can affect parameter estimation accuracy. Third, the samples' representativeness may be jeopardized. It may also make conducting a thorough study more challenging. As a result of any of these distortions, the validity of the trials may be jeopardized, and wrong conclusions may be made. Figure 2 illustrates the total number of null values in the data set for each attribute. Null values can be avoided by using simple statistical procedures such as the mean, median, and mode.



```
Total null values in Tenure 264 , its datatype float64
Total null values in WarehouseToHome 251 , its datatype float64
Total null values in HourSpendOnApp 255 , its datatype float64
Total null values in OrderAmountHikeFromlastYear 265 , its datatype float64
Total null values in CouponUsed 256 , its datatype float64
Total null values in OrderCount 258 , its datatype float64
Total null values in DaySinceLastOrder 307 , its datatype float64
```

Fig. 2. Null values in data

- Outliers, or values that deviate from the normal distribution of a set of variables, are another problem. Outliers are caused by a variety of factors, the most common of which are data entry and participant response issues. Outliers are data points that stand out from the rest of the distribution due to unusually high or low values. Outliers in the sample data can lead mean values to be underestimated or inflated, according to statistics. Outliers must be addressed before the data collection, which includes outliers, can be analysed. When the origins of outliers are determined, they are adjusted or replaced with substituted values. A distribution plot is used to detect the outliers. The left image in Figure 3 indicates the presence of outliers in the data. Outliers can be viewed in two ways: on top of or below the box. The percentile is then used to remove outliers from the data. There are no more outliers visible on the right side of figure 3.

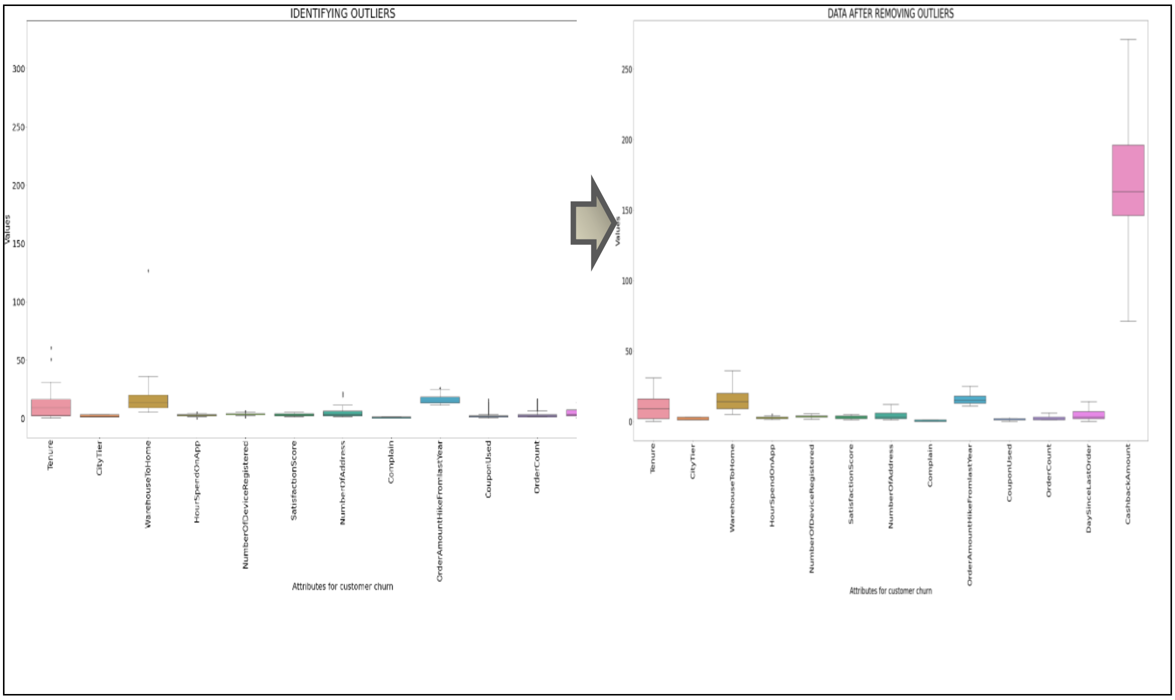


Fig. 3. Outlier detection and elimination using Bar plot

C. Feature Selection

Large dimensional data, particularly data with numerous attributes, is becoming more common in ML challenges. A huge number of experts rely on experimenting to address these challenges. Furthermore, these multidimensional variables and data can be used to extract significant traits. Statistical procedures were used to minimize noise and duplicated data. Do not, however, train a model with all of the features. Feature selection is crucial to develop the model with connected and non-redundant features [14]. Furthermore, it accelerates model training, decreases model complexity, clarifies model meaning, and improves metric performance. The selection of features is critical for four reasons. The first step is to reduce the number of parameters in the model by deleting extraneous features. Furthermore, to reduce training time, improve generalization, and avoid the curse of dimensionality, among other things. A dataset may contain various features or traits that influence the usability and applicability of the data in the field of data processing or analysis. When classifying, balance and imbalance data must also be considered. Furthermore, the goal is to create the best accurate model with the fewest errors feasible.

A variety of technologies can be employed to highlight the features of the dataset. The heatmap matrix is a popular method for visualizing the relationship between several characteristics. The association plot for the research mobile-commerce data is depicted in Figure 4.

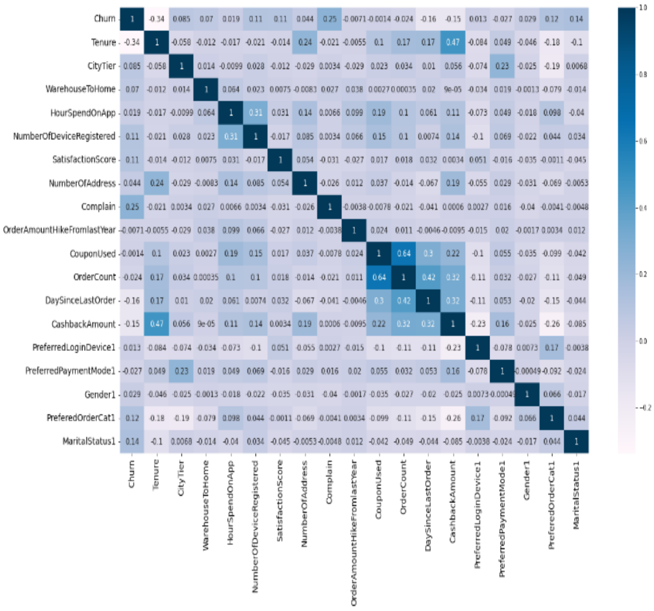


Fig. 4. Identifying most correlated features using the heatmap

A small number of features may be irrelevant in evaluating the prediction model. The feature with the lowest co-efficient value (near zero) should be eliminated from the feature list. This procedure is known as feature removal. Following the elimination steps, the data may show optimistic results. Figure 5 illustrates the eleven extracted features and one target variable from the raw data.

The total features of collected raw data from Kaggle are 19. The features will be reduced to 18, after removing customer ID, by using the basic understanding of the data. Then the correlate plot is employed to reduce the other irrelevant features. Now, the features are reduced to 11. The data features reduction at each step is shown in graphical format using figure 6.

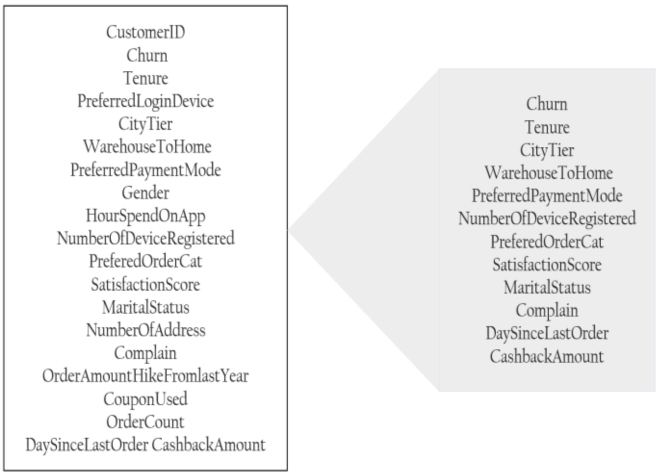


Fig. 5. Selection of most correlated features

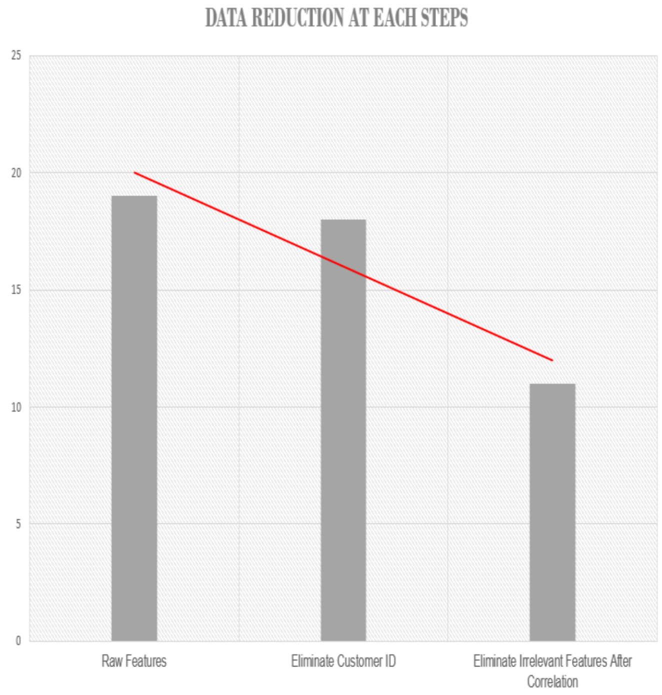


Fig. 6. Eliminating features at each step

D. Data Classification

Once the important data has been retrieved using feature extraction techniques, the classification method is utilized to predict which customers are going to churn and which are not utilizing the service. The classification methods that were employed in this study are detailed in detail below.

- SVM: SVM is a class of supervised learning techniques used for categorization [16]. They are part of a generalized linear classification family. SVM is unique in its ability to reduce empirical classification error while improving geometric margins. As a result, SVM is known as a Maximum Margin Classifier. SVM is used to map an input vector to a higher dimension where a maximum separation hyperplane is formed. On either side of the hyperplane, the data is separated into two sets of parallel hyperplanes. The separation hyperplane maximizes the gap between two adjacent hyperplanes. If there is a bigger margin or distance between such hyperplanes, it is assumed that they have smaller generalization mistakes. Figure 7 shows the SVM works. To defend against variables (attributes) with a wide range of values, scaling is required. To see training data in the form of a line, utilize the splitting (or separating) hyperplane.

$$w.x + b = 0 \qquad [1]$$

b is a one-dimensional vector, whereas w is a p-dimensional vector. Perpendicular to the w vector is the separating hyperplane. By including the offset parameter b, we may increase the margin. When b is missing, the hyperplane must pass through the origin, limiting the solution. We're interested in SVM and parallel hyperplanes because we're aiming for the highest possible margin. They can be described using the equation for parallel hyperplanes.

$$w.x - b \geq 1 \qquad [2]$$

$$w.x - b \leq -1 \qquad [3]$$

The distance between the hyperplanes is determined by geometry and the equation is $\frac{2}{|w|}$.

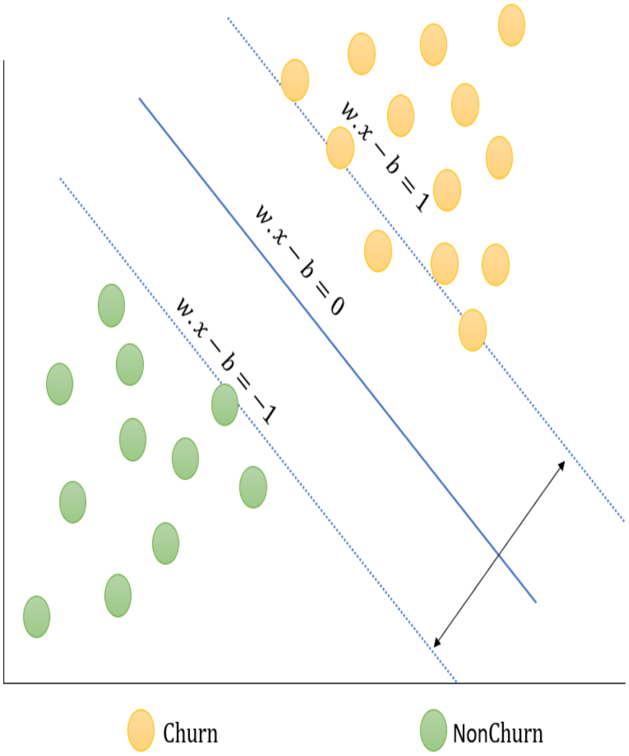


Fig. 7. SVM classification of Churn and non-churn

The confusion matrix obtained by SVM on mobile commerce customer churn test data is shown in figure 8. The 897 and 82 customers are predicted correctly as non-churn and churn by SVM. Then 42 and 105 customers are wrongly predicted as non-churn and churn.

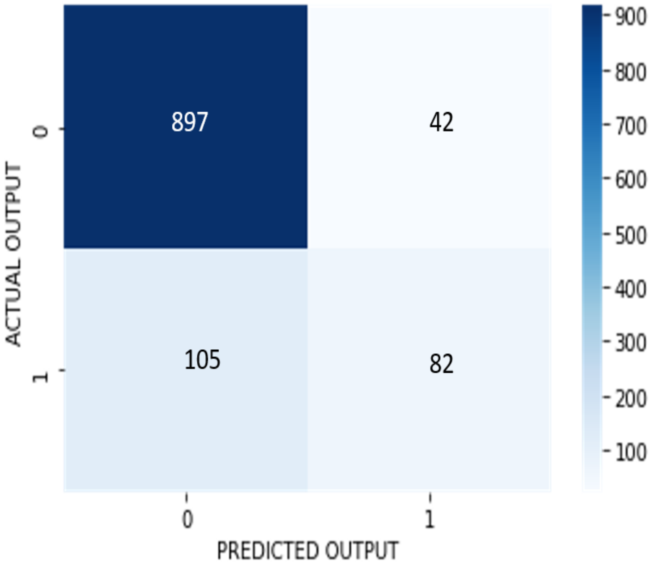


Fig. 8. SVM prediction on test data

- LDA: LDA can be used to simulate the differences between samples that have been classified into several categories. The purpose of the method is to optimize the variance ratio within and between groups. When this ratio is at its highest, the samples in each group have the least amount of scattering and the groups are the clearest. The goal is to maximize the below equation for a two-class discriminant problem where LDA's condition of equivalent group covariances is fulfilled.

$$S = \frac{pc_b p^T}{pc_w p^T} \qquad [4]$$

Where c_b and c_w are the covariance matrices between and within groups, and p is the direction in multivariate data space that best separates the two groups of samples. It is crucial to remember at this point that p is the eigenvector produced from the PCA decomposition of the matrices c_w and c_b. A multi-class discriminant problem can be generalized from the two-class discriminant problem. Because the LDA methodology is based on traditional estimators of location and covariance, it is sensitive to outlying samples, which implies that the method's performance degrades as the number of incorrectly allocated samples increases. Using robust estimates of data location and covariance instead of their traditional counterparts can assist in overcoming the LDA's lack of robustness. A robust pooled covariance can be calculated by employing several robust estimators and concepts. The journal [17] discusses several methods for making LDA more robust.

Figure 9 depicts the LDA-derived confusion matrix based on customer churn test data from mobile commerce. LDA correctly predicted non-churn and churn for customers 938 and 104. Then 24 and 60 consumers are mistakenly projected as non-churn and churn.

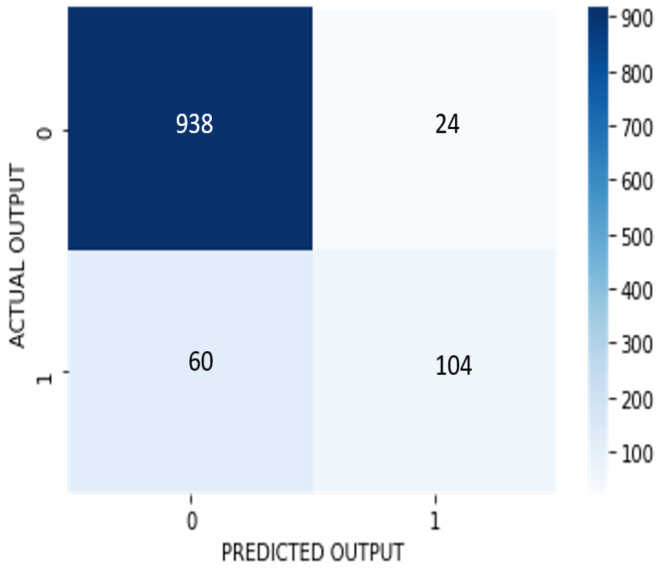


Fig. 9. LDA prediction on test data

- NB: The purpose is to construct a mechanism that will allocate a known categorization to the churn collection. With the use of this technology, consumer churn can be projected. NB is easy to construct since it does not necessitate the use of a complex recurrent parameterization technique. To put it another way, NB can be simply created for large data. Even if it isn't the ideal indicator for a certain goal, it may be trusted and works well [18]. Equations define the fundamental Bayesian theory.

$$P\left(\frac{W}{V}\right) = \frac{P(W)P\left(\frac{V}{W}\right)}{P(V)} \qquad [5]$$

Where:

V → Attributes

W→ Class

$P\left(\frac{W}{V}\right)$ → Probability of W given V

$P\left(\frac{V}{W}\right)$ → Probability of V given W

$P(V)$ → Probability of V

$P(W)$ → Probability of W

Hypothesis Maximum a Posteriori (HMAP) is used in Nave Bayes Predictor to increase the likelihood in each category.

$$H_{MAP} = argmax\ P\left(\frac{W}{v_1,v_2,v_3,\dots\dots,v_n}\right) \qquad [6]$$

$$= argmax\ P(W)\prod_{i=1}^{n} P\left(\frac{v_i}{W}\right) \qquad [7]$$

By utilizing the aforementioned equation, the data is categorized as to whether the client proceeds to churn or not. Figure 10 displays the NB-derived confusion matrix based on mobile commerce customer churn test data. NB accurately predicted 920 non-churn and 97 churn customers. Then, 25 and 84 consumers are incorrectly predicted as non-churn and churn, respectively.

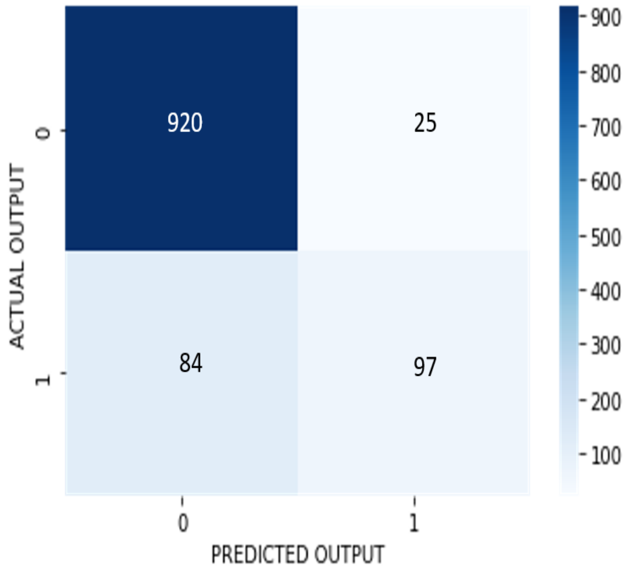


Fig. 10. NB prediction on test data

III. Discussion

Compare strategies for predicting customer churn in mobile commerce using the ML model are detailed in this section. The comparison provided for a thorough understanding of various prediction methods. The data was first collected from Kaggle, applied processing technique on raw data, employed an ML model to predict churn status. The status of the churn is binary whether Yes or No. A binary classifier model has exactly four generic outputs, as indicated by the confusion matrix in figures 8, 9, and 10. The first element in the matrix represents the True Negative (TN), the second one is False Negative (FN), then the bottom two represent the False and True Positive (FP & TP). The accuracy is calculated using the above-mentioned elements and the formula is

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad [8]$$

The accuracy of churn prediction by deploying SVM is 86.94%, likewise, the other two techniques LDA, and NB achieved the accuracy rate of 92.53% and 90.31%. The comparison, concludes that LDA is suitable for churn prediction on m-Commerce. The accuracy details of each model are given in Table II.

TABLE II. MODEL COMPARISON

Model	Accuracy
SVM	86.94%
LDA	92.53%
NB	90.31%

Table II is converted into a line graph and it is shown in figure 11. The peak of accuracy is obtained by the LDA method.

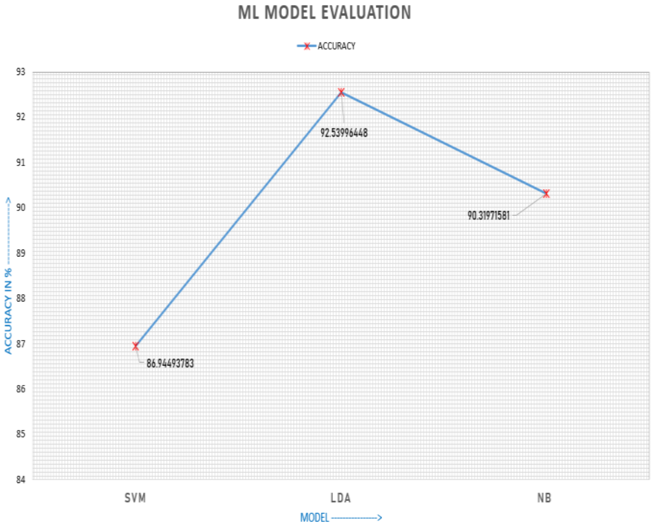


Fig. 11. Identifying best model using accuracy

IV. CONCLUSION

Forecasting m-commerce customer churn is a hot topic right now. Customer data from m-commerce is multidimensional and nonlinear. It is difficult to appropriately characterize changes in properties when automating churn prediction. The study presents an ML-based infrastructure that operates an end-to-end pipeline for predicting customer churn in m-commerce organizations. A study that compared different algorithms to find the best accurate model. This study allowed us to thoroughly investigate all 3 ML prediction methods. During the ML model prediction, learning about the factors to consider, among other things. After running all three models, it was discovered that the accuracy of LDA is superior to the other two. LDA predicted a total of 1042 correct churns from 1126 samples. It is intended to improve the analysis of customers who are projected to quit based on the reasons they cited for doing so in the future. Only those that voluntarily churn receive offers and rewards, resulting in hyper-targeted advertising. The churn forecast will be used in a mobile app. Customers who have been churned by the system will be notified about the latest deals and discounts. It will remind the customer to shop and notify them of any upgrades. It will alert the customer of any updates and prompt them to shop. It will give marketers a simple way to forecast churn using data and effectively carry out retention activities with their clients.

REFERENCES

[1]. Xiaojun Wu and Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-5, 2016.

[2]. Dick, A.S.; Basu, K. “Customer Loyalty: Toward an Integrated Conceptual Framework”, J. Acad. Marketing Science, vol. 22, pp. 99–113, 1994.

[3]. M Jaeyalakshmi, S Gnanavel, K S Guhapriya, S Harshini Phriyaa, K Kavya Sree, “Prediction of Customer Churn on e-Retailing”, International Journal of Recent Technology and Engineering (IJRTE), vol. 8 issue. 6, 2020

[4]. Gordini N, Veglio V, “Customers churn prediction and marketing retention strategies. An application of SVMs based on the AUC parameter-selection technique in B2B e-commerce industry”, Industrial Marketing Management, vol. 62, pp. 100-107, 2017

[5]. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," IEEE Access, vol. 7, pp. 60134-60149, 2019

[6]. Xiaojun Wu and Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-5, 2016.

[7]. Kaur and J. Kaur, "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 434-437, 2020

[8]. J. Latheef and S. Vineetha, "LSTM Model to Predict Customer Churn in Banking Sector with SMOTE Data Preprocessing," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), pp. 86-90, 2021.

[9]. C. Ju and F. Guo, "Research and Application of Customer Churn Analysis in Chain Retail Industry," 2008 International Symposium on Electronic Commerce and Security, pp. 670-673, 2008

[10]. S. Raeisi and H. Sajedi, "E-Commerce Customer Churn Prediction by Gradient Boosted Trees," 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 055-059, 2020.

[11]. X. Liu *et al*., "A Semi-Supervised and Inductive Embedding *Model for Churn Prediction of Large-Scale Mobile Games," 2018 IEEE International Conference on Data Mining (IC*DM), pp. 277-286, 2018.

[12]. Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., Prachayasittikul, V, “A practical overview of quantitative structure-activity relationship”, EXCLI Journal, vol. 8, pp. 74–88, 2009

[13]. Kwak SK, Kim JH, “Statistical data preparation: Management of missing values and outliers”, Korean J Anesthesiol, vol. 70, issue. 4, pp. 407-411, 2017.

[14]. Chen, Rung & Dewi, Christine & Huang, Su & Caraka, Rezzy., “Selecting critical features for data classification based on Machine Learning methods”, Journal of Big Data, vol. 7. pp. 26, 2020

[15]. Jaiswal JK, Samikannu R. “Application of random forest algorithm on feature subset selection and classification and regression”. In: World Congress on Computing and Communication Technologies, IEEE, pp. 65–8, 2017

[16]. Cortes, Corinna; Vladimir Vapnik, "Support-Vector Networks". Machine Learning, vol.20, issue. 3, pp. 273–297, 1995

[17]. Hubert M, Van Driessen K. “Fast and robust discriminant analysis”, Computational Statistics & Data Analysis, vol. 45, pp. 301–20, 2004

[18]. Wu, X., Kumar, V., Quinlan, J. R, Ghosh, J., Yang, Q., Motoda, H., & Philip, S. Y. “Top 10 algorithm in data mining”. Knowledge and information systems, vol. 14, issue. 1, pp. 1-37. 2008