Optimizing Customer Churn Prediction Models: A Comparative Study of SVM-Based Feature Reduction Techniques

1st Sagar Pathak
Department of Computer Science
University of Memphis
Memphis, TN
spathak1@memphis.edu

2nd Kisan Thapa
Department of Computer Science
University of Massachusetts
Boston, MA
kisan.thapa@umb.edu

3rd Prabin Sharma
Department of Computer Science
University of Massachusetts
Boston, MA
prabin.sharma@umb.edu

***Abstract*—Customer churn risk, the likelihood of customers discontinuing their relationship with a business, poses a significant challenge to companies across industries. Factors such as customer satisfaction, market dynamics, and lack of engagement contribute to churn risk, impacting revenue and long-term profitability. This paper addresses the challenge of analyzing and predicting customer churn risk, employing machine learning techniques to classify customers into high, medium, and low churn risk categories. Three Support Vector Machine (SVM) models were developed, utilizing different feature reduction methods including Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) with an optimal component selection strategy. Our experimental results demonstrate that the SVM model employing LDA outperforms others, achieving an accuracy of 93.47% and an F1 score of 0.93 on the test dataset. This research underscores the effectiveness of machine learning in mitigating customer churn risk, offering valuable insights for businesses to enhance customer retention strategies and sustain long-term success in a dynamic market environment.**

***Index Terms*—Customer churn risk, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA)**

I. Introduction

In this era of technological advancements, businesses play an integral role in connecting everything. Even a simple cup of tea involves a complex network of businesses that engage in tea farming, sugarcane farming, shipping, and processing of raw materials, followed by retailing through supermarkets and groceries. The market is vast, with a wide range of products catering to different niches and target customers. The success of a business hinges on its ability to remain profitable, which can only happen if the product has enough customers to support it. Customer churn, or the loss of customers, can be detrimental to a company's growth and success. Therefore, it is essential for businesses to focus on retaining their customers. There are several factors that are related to customer churn, and it also depends on the business and type of customers. Therefore, analyzing customer churn risk is a challenging task that requires careful evaluation of each property correlated with the customers. Machine learning approaches are handy tools for conducting such analyses and can simplify the intricate nature of this problem. This assignment is a sample project that aims to classify each customer into three categories of churn risk: high, medium, and low.

II. Problem Definition

The process of categorizing customer churn risk is difficult due to the intricate nature of the business they are associated with, as well as the impact of numerous other factors. Churn analytics is a method used to measure and understand the frequency and location of customer attrition, while also identifying features and functions that may improve customer retention. This type of analysis is vital for obtaining an overall performance summary, identifying areas for improvement, and determining which channels provide the most value. The primary aim of churn risk analysis is to increase business profitability. However, identifying relevant attributes related to customer happiness, satisfaction, dissatisfaction, and other factors can be challenging, and such attributes can vary from one customer to another due to human dynamics. However, using the power of machine learning and data analytics, a robust prediction model can be developed by understanding the complex interplay of these factors in historical data. The complexity of this issue is extensive, and machine learning technology can be helpful in conducting churn analysis.

III. Background Study and Related Works

There are several related works on churn prediction in various business areas. For instance, Zhao et al. (2005) [1] conducted research on customer churn prediction using an improved one-class support vector machine. They found that the Gaussian kernel support vector machine performed the best with an accuracy of 87.15%. The researchers utilized a wireless industry customer churn dataset, and due to the small number of negative examples, they introduced an improved one-class SVM. Similarly, Xia et al. (2008) [2] worked on improving the prediction abilities of machine learning methods using support vector machines. They applied structural risk minimization and predicted cases for home

and foreign carriers, with a dataset containing 3333 instances and 21 features. Their findings revealed that the radial basis kernel support vector machine performed the best, with an accuracy of 90.88%. In another study, Shaaban et al. (2012) [3] discovered that support vector machines had the highest accuracy, with 87.3%, in a dataset containing 5000 instances with 23 attributes. They predicted and clustered the dataset into three categories of customer churn and compared the results with Decision Tree and Neural Network models.

Zhou et al. (2019) [4] conducted research on a combined scheme for predicting appetency and up-selling in the telecommunications industry using support vector machines with two different kernel functions. This work directly correlates with the customer churn risk analysis. The polynomial kernel function yielded the highest accuracy of 91%. The proposed method involves utilizing the posterior probability of customers to determine if non-churning customers will either buy a new product or upgrade. The polynomial kernel function was found to have the lowest error rate and is thus deemed the optimal model for predicting appetency and up-selling in the telecommunications industry. Furthermore, using various sample sizes demonstrated that a range of sample sizes is equally effective in predicting the model's performance.

IV. Objective

The study aims to solve the problem of churn risk prediction by utilizing customer data that consists of 2066 data points and 15 relevant features (excluding ID and CHURNRISK). To achieve this, the Radial Basis Kernel Support Vector Machine (Kernel-SVM) is used in this project. Furthermore, the objective of the assignment is to gain knowledge and implement Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to reduce the features' dimensionality and apply SVM on the lower dimensional feature space. The study also aims to deepen the understanding of different model evaluation and performance metrics, including but not limited to the confusion matrix, precision, recall, F1-score, and ROC curve. By employing these evaluation metrics, the performance of the model is assessed accurately. This study is significant as it can help businesses in some industries to accurately predict churn risk, leading to better customer retention strategies and improved customer satisfaction.

V. Methodology

A. Data Preprocessing

The goal of data preprocessing is to make the data suitable for further analysis, modeling, and interpretation and to remove any biases or inaccuracies that may impact the results. We can use exploratory data analysis (EDA) to discover patterns and distinguish anomalies. Heatmap can be useful because it allows us to visually represent and explore complex relationships between variables in a data set. It is possible to feed the data directly into the SVM model but for better accuracy, EDA is performed and removed features that do not contribute to the overall accuracy of the model.

The initial size and dimension of the housing price dataset is 35122 and 2066x17 respectively, where the first column is for the ID attribute. There are 15 independent variables and the target variable named as "CHURNRISK". As the data values were preprocessed, few steps were taken beforehand, which were removing the leading or trailing spaces from the dataset and dropping rows which has null values and also, performing the data type exploration on the dataset. The customer dataset contains 12 numerical feature columns and 4 categorical feature columns.

1) Exploratory Data Analysis (EDA): Exploratory Data Analysis (EDA) is a method of studying and analyzing data sets to identify important characteristics and relationships. This is done by using various statistical and visual techniques to gain insights into the data. EDA is usually performed at the start of a data analysis project to understand the data better and identify any issues that need to be addressed. In this project, the customer dataset has numerical and categorical variables, so the numerical feature exploration and categorical feature exploration was conducted as a part of exploratory data analysis.

2) Numerical Feature Exploration: Numerical feature exploration helps to analyze and understand the numerical variables in the dataset. Heatmap is one of the tool to explore on numerical features. It helps to display correlation matrices, where each cell in the matrix represents the correlation coefficient between two variables. The purpose of the heatmap is to find out the variables which have a high correlation to each other and analyze the multi-colinearity as well. Multicolinearity occurs when multiple features in a statistical model are highly correlated with each other. This undermines the model's validity as the two variables are not independent of each other, and the model can't accurately determine the significance of each in predicting the target outcome. Below is the table which describes the correlation and its level of strength.

TABLE I
TABLE DESCRIBING THE CORRELATION AND IT'S LEVEL OF STRENGTH.

r Value	***Strength***
r = ±1	Perfectly Correlated
$\pm0.7 \leq r < \pm1$	Strongly Correlated
$\pm0.4 \leq r < \pm0.7$	Moderately Correlated
$\pm0.1 \leq r < \pm0.4$	Weakly Correlated
$0 \leq r < \pm0.1$	Negligible Correlation

This table will be used to identify strongly correlated features and proceed with the data-cleaning process, where we remove features that do not contribute much to the model performance.

3) Categorical Feature Exploration: This is a process of analyzing and understanding categorical variables in a dataset. Categorical features are variables that represent discrete values, such as colors, labels, or categories. Categorical feature exploration can be done using techniques such as Chi-squared

test, frequency tables, bar charts, pie charts, and stacked bar charts. The project uses the bar chart to analyze the categorical features.

B. Data Encoding

This process converts categorical data into a numerical representation that can be used as input for machine learning algorithms. This is necessary because many machine learning algorithms can only process numerical data. Common techniques for data encoding in machine learning include one-hot encoding, label encoding, and binary encoding. The choice of encoding technique depends on the nature of the data and the requirements of the machine learning algorithm being used. Since we have three different class values for CHURNRISK, label encoding is used on this dataset.

1) Label Encoding: We will perform label encoding for this experiment because it is simple and effective, where each category is represented as a numerical value. The "High" risk is labeled as 0, the "Medium" risk is labeled as 2, and the "Low" risk is labeled as 1 on this experiment.

C. Data Preparation

The numerical and categorical data are merged after the preprocessing using exploratory data analysis, data encoding techniques, and data cleaning. The data standardization on the numerical features is performed using StandardScaler. And then the train/test split is applied into the merged dataset in such a way that 80 percent resides as the training data, whereas the remaining 20 percent will be used as the testing set.

After the split, the customer dataset had 1652 datapoints in the training set and 414 datapoints in testing set. The dataset is then became ready and modeled using Radial Basis Kernel Support Vector Machine (Kernel-SVM). Classifier performance is assessed using various evaluation metrics, such as confusion matrix, precision, recall, F1-score, and ROC curve. Additionally, the computational times for training and testing phases are also evaluated to determine the classifier's efficiency.

VI. Model Description

Radial Basis Kernel-Support Vector Machine (Kernel-SVM) is applied to the customer churn prediction dataset and also the dataset is processed using feature reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and different versions of models are generated.

A. Radial Basis Kernel-Support Vector Machine (RBF-SVM)

The Radial Basis Kernel-Support Vector Machine (RBF-SVM) is a highly effective machine learning algorithm for classification and regression tasks. It falls under the category of kernel methods and uses the radial basis function (RBF) kernel, which calculates similarity between two data points based on their Euclidean distance and a hyperparameter called gamma. The algorithm identifies a hyperplane that maximally separates the data points with a high margin in the high-dimensional space, using the support vectors that determine the location and orientation of the hyperplane to classify new data points. RBF-SVM is a powerful algorithm that can handle non-linearly separable data and achieve high accuracy in various applications, such as image classification, speech recognition, and bioinformatics, but requires careful tuning of hyperparameters such as gamma and the regularization parameter for optimal performance. The gamma parameter determines the range of influence of a single training example, where low values indicate a wide range and high values indicate a narrow range. Essentially, gamma acts as the inverse of the radius of influence of the model's selected support vectors. When gamma is set to a high value, the support vector's area of influence is so small that no amount of regularization with can prevent overfitting. So I am keeping low gamma values for this research.

B. Principal Component Analysis (PCA)

PCA is a statistical method that identifies patterns in complex data sets by identifying the directions with the highest variance and projecting the data onto these directions. The resulting principal components are new variables that are uncorrelated and capture the majority of the variability in the original data. By reducing the dimensionality of the data, PCA can simplify analysis, visualization, and computation, and it has numerous applications in fields such as machine learning, signal processing, image analysis, and finance.

C. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a statistical method that is commonly used for classification and reducing the number of variables in a dataset. The primary objective of LDA is to determine the best linear combination of variables that can differentiate between two or more classes of data. To achieve this, LDA aims to maximize the distance between the means of different classes, while minimizing the variation within each class. This is accomplished by computing the between-class and within-class scatter matrices and then calculating the eigenvectors of their ratio matrix. The eigenvectors with the largest eigenvalues represent the directions in which the data can be most effectively separated, and they are used to project the data onto a lower-dimensional space.

VII. Model Evaluation & Performance Metrics

A. Confusion Matrix

The confusion matrix is a tabular representation used to evaluate the classification model's performance by comparing the data's predicted and true class labels. It displays the correct and incorrect classifications of each class in the data. Accurate classifications are shown in the diagonal of the matrix and the rest of the other matrix elements show the incorrect classifications. The confusion matrix provides various performance metrics like accuracy, recall, precision, and F1-score, which help assess the classification model's overall performance.

B. Precision and Recall

Precision and recall are commonly used performance measures to assess how well a classification model performs. Precision assesses the accuracy of positive predictions by calculating the ratio of true positives to the sum of true positives and false positives. In essence, precision gauges how many of the positive predictions are correct. A high precision score implies a low rate of false positives. Recall, on the other hand, also referred to as sensitivity or true positive rate, assesses how well the model identifies positive instances among all actual positives. This is calculated by dividing the number of true positives by the sum of true positives and false negatives. Recall evaluates how many of the actual positive instances are identified as positive. A high recall score implies a low rate of false negatives.

C. F1-Score

The F1 score is one of the best evaluation metrics for classification systems, as it combines precision and recall. Precision is the ratio of true positives to the total number of positive predictions. A lower false positive rate results in higher precision, while recall measures sensitivity or the true positive rate. Both metrics are important, and the F1 score combines them using their harmonic mean. A higher F1 score indicates both high precision and recall, making it a reliable measure of model performance.

D. ROC Curve

The ROC curve is a graphical plot used in binary classification models to show the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various classification thresholds. By plotting TPR (sensitivity) against FPR (1 - specificity) at different threshold settings, the area under the ROC curve (AUC) is calculated, which is a measure of the overall performance of the model. A value of 1 represents a perfect classifier, while a value of 0.5 represents a random classifier. The ROC curve is especially useful in machine learning for evaluating models in situations where class imbalance exists. It is a valuable tool for comparing different models as it provides a single score that summarizes performance across different threshold settings. The target variable we have here in this dataset has three different values, High, Low and Medium. So we cannot implement the ROC Curve directly thus this project uses OvR (One vs Rest) strategy.
In this project, all of these evaluation metrics functions were imported from the scikit-learn package and used for model evaluation except for a few functions such as for ROC Curve.

VIII. Experiment and Results

The experiment is conducted on Google Colab using the configuration of Intel Xeon CPU @2.20 GHz with 12 GB GDDR5 VRAM. I analyzed the performance of the Radial Basis Kernel-Support Vector Machine (RBF-SVM) model using different evaluation metrics and in different setting, one with feature reduction through Exploratory Data Analysis (EDA), the second with Principal Component Analysis and the third with Linear Discriminant Analysis (LDA). In the beginning, the customer dataset is loaded which has 2066 rows and 17 columns. The target variable in this dataset is "CHURNRISK" and there is one "ID" column that is removed as a part of data cleaning since it does not contribute much to the data analysis. As a part of data preprocessing, the leading or trailing spaces from the data values were removed. The null values were also removed from the dataset.
The exploratory data analysis is conducted on numerical and categorical feature variables. There were 5 integer features variables and 7 float variables summing up 12 numerical feature variables. Numerical feature exploration is conducted by plotting the heatmap, which helps to identify the multi-collinearity of the features in the dataset.

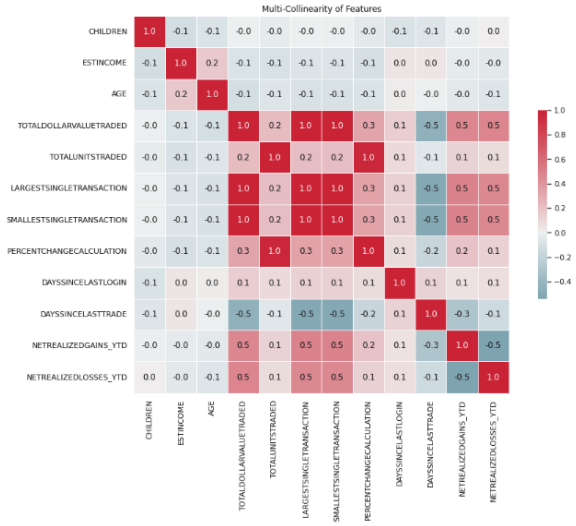

Fig. 1. Multi-collinearity of customer features (Heat map)

As we can see on the heatmap above (Figure 1), there are features which are strongly correlated with other features such as, TOTALUNITSTRADED and PERCENTAGECHANGECALCULATION, TOTALDOLLARVALUETRADED and SMALLESTSINGLETRANSACTION and SMALLESTSINGLETRANSACTION. We should drop these for the better modal prediction which we will do in data-cleaning process.
These categorical features need data encoding before fitting into the regression model. Thus, label hot encoding is used to replace these feature values with 0, 1, and 2. The "High" risk is labeled as 0, the "Medium" risk is labeled as 2, and the "Low" risk is labeled as 1 in this experiment. A similar replacement is made for other categorical attributes.
The numerical and categorical data are then merged to form a combined dataset. The training and testing dataset is then split in the fashion of 80% training and 20% testing samples. After the split, there are 1652 rows in the training dataset and 414 data points in the testing dataset.

The following is the heat map plot and the distribution

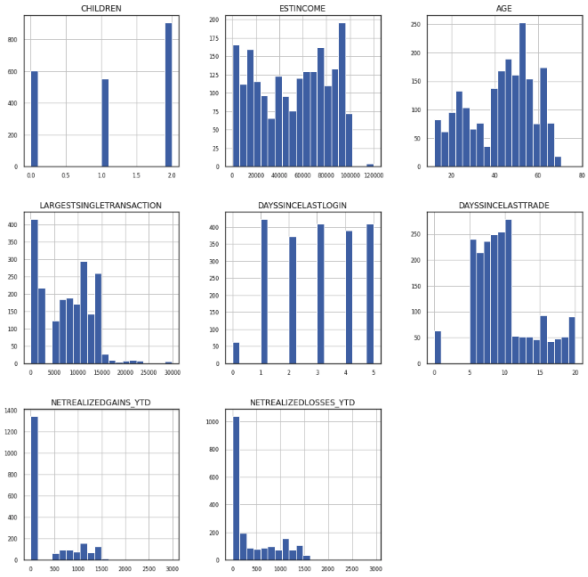

Fig. 2. Descriptive statistics of categorical features

plot for the feature columns. The distribution plot is the bar graphics where the frequency with its value is plotted.

The Radial Basis Kernel Support Vector Machine then built using the training dataset for the customer dataset. So there are basically three different SVM models. One is using created after the EDA process, the other is processed after the PCA analysis and the final model is generated with the help of LDA feature reduction technique. These models are evaluated using different metrics.

The confusion matrix for the model built through the EDA analysis is shown below

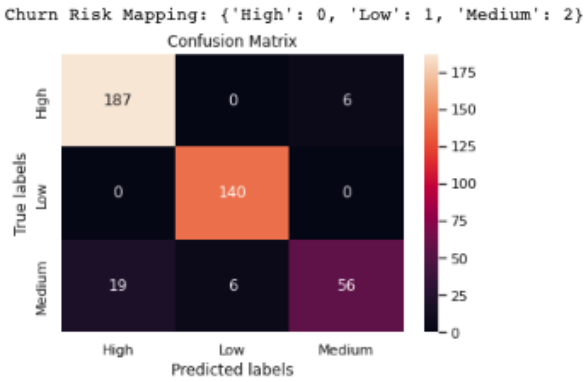

Fig. 3. Confusion Matrix 1

The overall accuracy of the model built through the EDA analysis in task 1 was 92.51 percentage and the overall F1 score is 0.93. The Receiver Operating Characteristic (ROC) curve for this model is shown below.

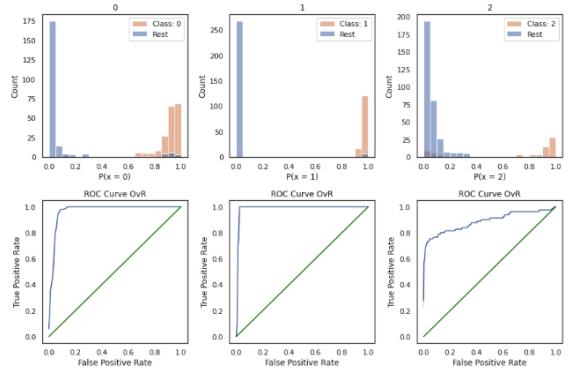

Fig. 4. ROC Curve using OvR(One vs Rest) strategy for model 1

The label in graphs, 0, 1, and 2 represents ”High”, ”Low” and ”Medium” customer churn risk respectively.

Principal Component Analyis (PCA) is done for feature reduction and then applied the SVM on both lower dimensional feature space. PCA needs a suitable number of component k and this can be selected using different techniques. I used explained variance technique to figure out the number of components required while downsizing the features. The figure below shows the explained variance for PCA which helps us to determine the number of columns that can be used for PCA.

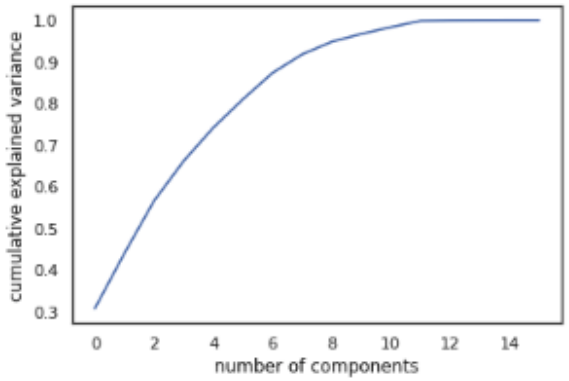

Fig. 5. Explained Variance Analysis on customer dataset

As we can see on the graph above, the first 5 components contain about 85% of the variance and the first 8 components contain about 95% of the variance while we need altogether 10 components to describe 100% of the variance. Now, let’s do the PCA using 5 components. After the PCA operation, the new dataset is reshaped to 1652x5. And then a similar SVM model is generated with this dataset. The following is the confusion matrix that I got for the second model (After PCA).

The overall accuracy of the model built through the PCA was 66 percent and the overall F1 score is 0.66. The Receiver Operating Characteristic (ROC) curve for this model is shown below (Fig. 7).

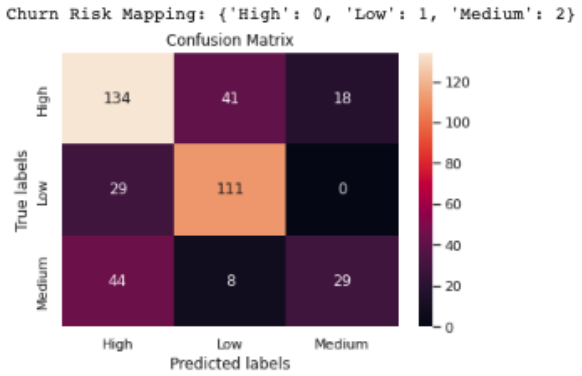

Fig. 6. Confusion Matrix for the model after PCA

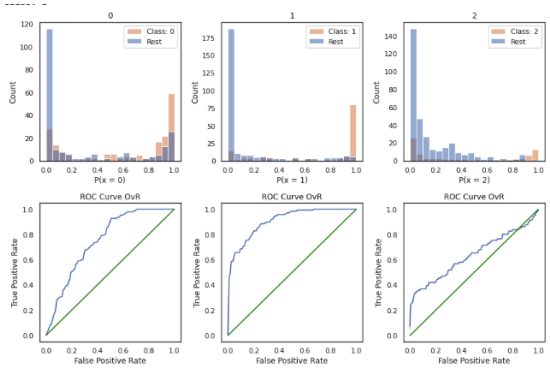

Fig. 7. ROC Curve using OvR(One vs Rest) strategy for model 2

Then the final model is created using the Linear Discriminant Analysis (LDA). The optimal number of components are selected using explained variances, and we got only 2 components for LDA analysis. The third model is built. The following is the confusion matrix.

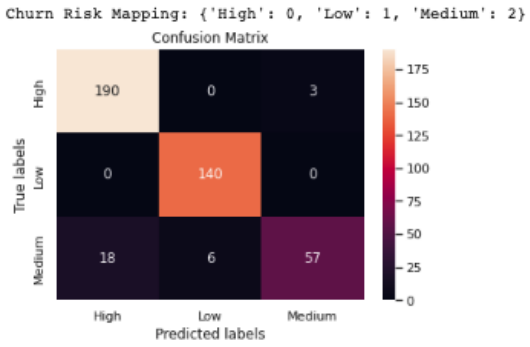

Fig. 8. Confusion Matrix for the model after LDA

The overall accuracy of the model built through the LDA was 93.47 percent and the overall F1 score is 0.93. The Receiver Operating Characteristic (ROC) curve for this model is shown below.

IX. Discussion

The comparison of three models built on a dataset after performing some feature reduction techniques such as Exploratory Data Analysis with numerical and categorical feature analysis. The aim of these techniques is to reduce the dimensionality of the dataset and select the most important features that contribute to the target variable. After applying the feature reduction techniques, the dataset was standardized using StandardScaler to make it more scaled. StandardScaler is a common technique used to scale the data and bring it to a standard normal distribution with a mean of zero and a standard deviation of one. Then, Principal Component Analysis (PCA) was performed to further reduce the dimensionality of the dataset. PCA is a popular technique used for feature extraction and dimensionality reduction, which transforms the original features into a new set of orthogonal features called principal components. These principal components are selected based on their ability to explain the maximum variance in the data. Two models were built using the transformed dataset obtained after PCA. However, the performance of the second model was not good compared to the other models, indicating that the selected principal components were not able to explain the variation in the target variable. Finally, a third model was built after performing Linear Discriminant Analysis (LDA), which is a supervised dimensionality reduction technique that aims to maximize the separation between different classes in the data. The third model performed the best with an accuracy of 93.47% and an F1 score of 0.93, indicating that the selected features using LDA were able to explain the maximum variation in the target variable and provide better separation between different classes in the data.

X. Conclusion

Three different SVM models were created using different feature reduction techniques such as Exploratory Data Analysis (EDA) with numerical and categorical analysis, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) with an optimal component selection strategy. Among these models, the one that employed LDA performed the best on the test dataset, achieving an accuracy of 93.47% and an F1 score of 0.93.

References

[1] Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Customer churn prediction using improved one-class support vector machine," in Advanced Data Mining and Applications (X. Li, S. Wang, and Z. Y. Dong, eds.), (Berlin, Heidelberg), pp. 300–306, Springer Berlin Heidelberg, 2005

[2] W.-d. J. Guo-en XIA, "Model of customer churn prediction on support vector machine - sciencedirect." https://www.sciencedirect.com/science/article/abs/pii/S187486510960003X. (Accessed on 03/02/2023).

[3] K. H. e. a. Essam Shaaban, "A proposed churn prediction model." https://www.researchgate. net/publication/236625937 A Proposed Churn Prediction Model. (Accessed on 03/02/2023).

[4] L.-Y. Z. et al., "Combined appetency and upselling prediction scheme in telecommunication sector using support vector machines." https://www.mecs-press.org/ijmecs/ijmecs-v11-n6/ IJMECS-V11-N6-1.pdf. (Accessed on 03/02/2023)