Customer Churn Prediction using ML models

1st Shiny Shamma Kota
MS in AI (11698716)
UNT
ShinyShammaKota@my.unt.edu

2nd Sai Bhavani Shankar Chintapalli
MS in CS (11724519)
UNT
saibhavanishankarchintapalli@my.unt.edu

3rd Hanu Vamsi Putta
MS in CS (11704879)
UNT
hanuvamsiputta@my.unt.edu

4th Sreenivas Varma Chamarthi
MS in DE (11725404)
UNT
sreenivasvarmachamarthi@my.unt.edu

***Abstract*—Prediction of customer retention of a company subscription service or product; in other words, churn prediction is a desirable study both by the corporate and the academia. Machine learning models are developed to the extent that one can predict future events by proper model training. The customer Churn dataset is subjected to logistic regression model, Gradient Boosting, Random Forest, and Neural Network model. The data set features are reduced in dimension using PCA. This dataset is also subjected to these four machine-learning models. Finally, based on different performance metrics, this project aims to give a best-fit model.**

***Index Terms*—EDA, PCA, Machine Learning Models**

I. Introduction

1) Motivation: In a world driven by capitalism, the income and profit of a company have significant importance on the company's survival. As most of the world's businesses depend on customers, it is vital for the company to have a stream of faithful customers adhere to buying their products, and it is more cost-effective to retain old customers than to procure new ones in this type of environment [1]. This very nature of retaining old customers, which is an environmentally optimal decision, became the source of our motivation to take up this project.

2) Significance: Customer churn prediction is significant to the corporate businesses as it gives an insight into the data for better decision making by the companies. In doing so these businesses can develop strategies to retain their existing customers and prevent dip in the revenue. Customer churn knowledge will also help businesses to increase the efficiency in customer service and optimize their operations. This in turn gives the company a competitive edge, and address their shortcomings. With the aid of machine learning models, it gives an increased probability to predict future events, in this case customer churn.

3) Objectives: The primary goal of this project is to develop a machine learning model which can predict customer churn with the highest accuracy with any company's customer churn data set. Secondary goals in doing this project are:

- Understand the dataset and perform data analysis.
- Perform Dimension Reduction using PCA-Principal component analysis. This is done to check whether dimension reduction betters the performance or decreases it.
- Find the performance of each model on the dataset before PCA and after PCA.
- Propose a model based on the performance metrics.

II. Design and Features

A. Literature Review (Related Work)

Churn rate is the percentage of customers that used to use a company's product or a service and stopped using it. For a company to have increased volume in customers it's growth rate must exceed churn rate [2]. For a company to have increased volume in customers it's growth rate must exceed churn rate. A decrease in the churn rate by 5% will increase the company's earnings by 25 to 85% as observed by [3]. This makes churn rate prediction desirable for research. Many studies are done in order to predict the churn rate on the company's client information. In the study by [4], proposed deep learning models and machine learning models, and analysis on the result showed that deep neural network models gave better accuracy compared to machine learning models. A combination model of decision Tree and Neural network was proposed by [5] to predict customer churn rate. The result showed that it has higher accuracy compared to single machine learning models. An average accuracy of 81% percent is observed when ensemble neural network based classifiers are used for churn prediction [6]. An experimental study by [7] shows that the PCA-based prediction method is potential way to forecast, because the algorithm is simple, and it can reduce dimensions of data and simplify data processing. These studies have laid a background for our study and selection of ML models.

B. Dataset

The Telco customer churn data set used for this project has been downloaded from kaggle [8]. The data set has information related to telecommunication companies, namely Telco and their customer's details. The features included in the data set are customer demographics such as gender, whether the person is a senior citizen or not, whether they have a partner, and whether they have any dependents on them. The data set also has information about the services availed by the

customer related to the company. (7043, 21) is the dimension of the dataset taken for our analysis.

C. Detail design of Features

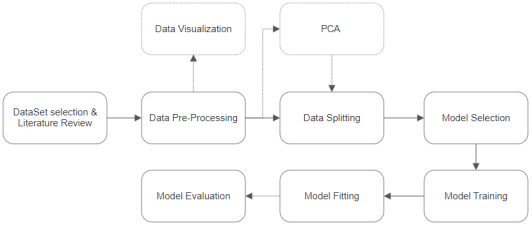

Fig. 1: Flow Chart of the project

1) Data preprocessing: With the help of a programming language, namely jupyter notebook, The data is inspected for any missing values or null values also, any feature that is of no use to prediction modeling is dropped, and the categorical string values in the data set are encoded into numerical forms for further use in Machine learning modeling.

2) Data Visualization: The column 'churn' becomes our dependent variable, and all the remaining columns become independent variables. Depending on the requirement, we perform different data visualizations for our data set; we have chosen pie charts, count plots, and distribution plots to visualize the information the data set represents.

3) Dimension Reduction: . Our data set has features amounting to 19 independent variables, which is more significant in number; therefore, we perform PCA to reduce the dimension of the data set. With the help of SPSS we perform Dimension reduction, which will act as a comparative data set in analyzing the existing data set.

4) Model Selection: Churn prediction is a classification model that predicts whether the customer will opt out of the company's services. Based on the background study, this project has adopted machine learning models, namely logistic regression, gradient boosting, random forest, and neural network, to perform our prediction analysis on the data set.

5) Model Training and prediction: The encoded data set is split into training and testing data sets. The Dimension reduced data set is also split into the training and testing data sets for machine learning prediction modeling. Model fitting is done on the training data set, and using the testing data set, we predict the dependent variable for the corresponding machine learning model.

6) Performance Metrics: The predicted dependent values are compared with the existing dependent variables, and the model performance is measured using its accuracy, precision, f1 score, recall, and confusion matrix.

7) Features of the project:

- This project employs four machine learning models, and their exhaustive predictive capabilities through their accuracy scores are examined to conclude.
- This project is unique for its comparative studies of the machine learning models of the data before and after performing Principal Component Analysis.

III. EDA- Exploratory Data Analysis (Analysis)

The study made use of dataset [8], which contains details about customer data and Telco. Some of the features in the data set are customer personal information such as gender, whether or not the person is a senior citizen, whether or not they have a partner, and telecom-related information like whether or not the person uses paperless billing or electronic payments. Details regarding the services the client utilized that were offered by the company are also included in the data set.

A. Descriptive Statistics

In [8] dataset, there are two numeric data features: tenure and monthly charges. The average Monthly charges came around to be 64.76. This implies that every subscribed person pay 64.76 currency for the company services The average Tenure of the customers in the company : 32.37 months. This implies that company has held onto an average customer loyalty for 32.37

B. Data Visualization

1) Demographics: The percentage of females: 50.48%

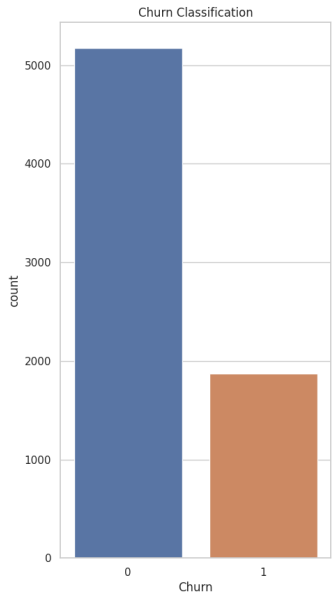

Fig. 2: Count plot of the churn feature

Figure 2 shows the visualization of last column i.e. Churn, which shows if customer opted out of their service or not. The dataset has considered equal ratio of male and female customers.

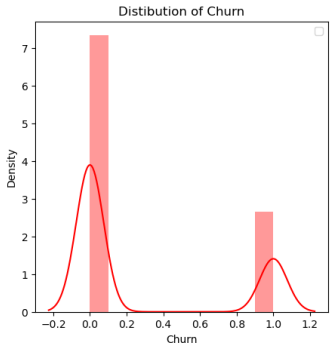

Fig. 3: Density plot of churn

The churn graph is now shown in density and churn as in fig 3 instead of Count and Churn.

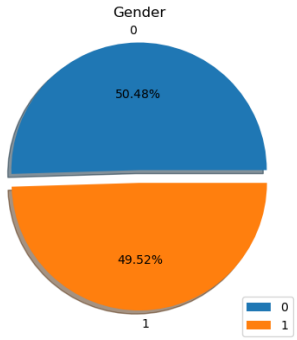

Fig. 4: Gender representation of customers

The Figure 4 shows a pie chart which represents how much percentage of each gender that the telecom service provider associated with. Approximately the genders are distributed equally.

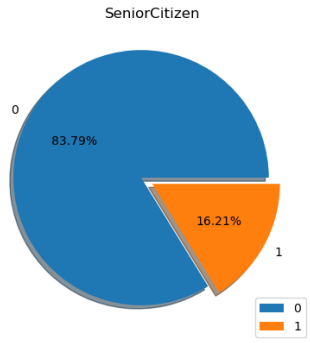

Fig. 5: Seniority representation of customers

The Figure 5 shows a pie chart describing how much percentage of the customers are senior citizens, Only 2/10th are seniors that means that the maximum number of customers that avail services from these companies are not elderly.

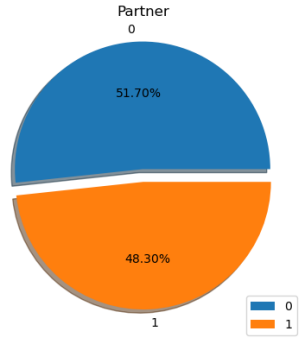

Fig. 6: Percentage of customers with partners

The Figure 6 is a pie chart which shows the percentages of customers with partner and without partner. From the graph we can say that approximately half of the customers do have a partner.

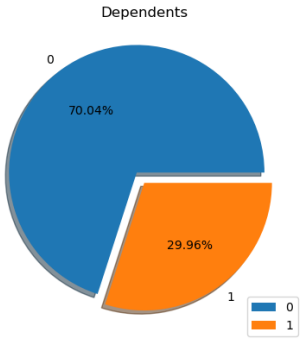

Fig. 7: Percentage of customers with dependents

The Figure 7 is a pie chart which shows the percentages of customers who have dependents and who do not have any dependants with them. From the graph, approximately more than 3/4th of customers have dependants with them.

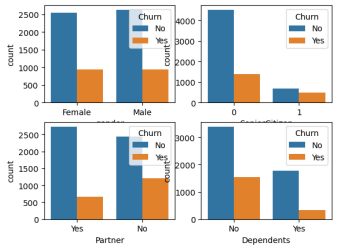

Fig. 8: four count plots of demographic data in data-set

The Figure 8 shows 4 different bar graphs consisting of all the above mentioned categories. Each graph represents how many customers are opted out and how many customers have remained using their services in each of the category.Majority of the customers who have partners,the customers who are not senior citizens and the ones with dependants, all continued to use their services.

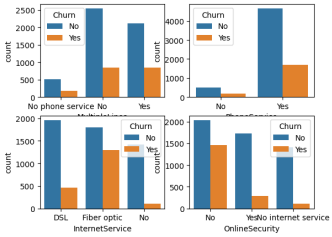

Fig. 9: Count plot of next four features

The Figure 9 also shows 4 different bar graphs from which we can observe, how many customers have opted out or remained using the services in terms of the service.The four different services considered are PhoneService, MultipleLines, Internet Service, OnlineSecurity.

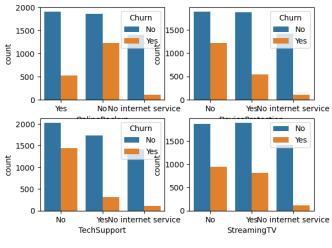

Fig. 10: Count plot of next four features

The Figure 10 also shows the 4 different bar graphs from the categories Online Backup, Device Protection, Tech Support, and StreamingTV. It also shows whether the customer has opted out of their service or stayed with them.

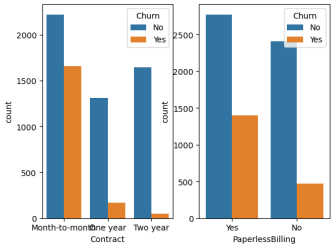

Fig. 11: Count plot of next two features

The Figure 11 show 2 graphs from categories Contract and Paperless Billing. It also shows whether the customer has opted out of their service or stayed with them.

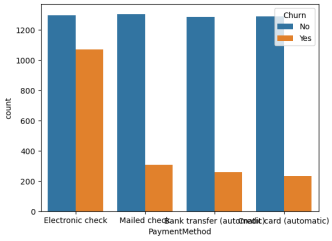

Fig. 12: count plot of payment method

The Figure 12 shows a graph specifying the customers who stayed with their previous payment plans and who opted out. Here the customers who use electronic checks as their payment method opted out, while the people using automatic payment plan have opted to stay.

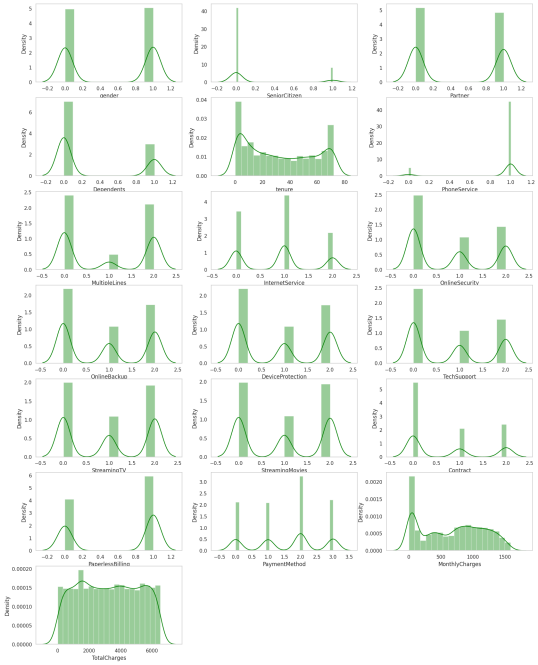

Fig. 13: Density Distribution plot of independent and dependent variables

The Figure 13 takes all the categories mentioned in above graphs and plots them using density instead of count. This gives the plot a curve in bar graph plot.

IV. Implementation

We have used Google colab (jupyter notebook) and SPSS software to reach our goal.

A. SPSS

In IBM SPSS perform dimension reduction (PCA) on the encoded data set. The steps involved are:

- Load the encoded dataset.
- In 'Analyze,' select dimension reduction.
- Further select 'Factor...'.
- Select all the independent features and transfer them into the variables column.
- In 'Extraction' select Scree Plot as per 14.
- Other features to be selected are: In descriptions, select the initial solution and KMO and Bartlett's test of sphericity. In scores, select regression. In options, place the absolute value below 0.1.
- Click on okay to generate output.

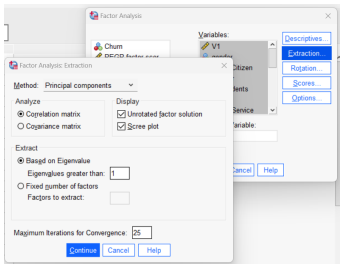

Fig. 14: Scree Plot computation using SPSS.

The Figure 14 shows Scree Plot compution using SPSS. It displays the configuration they used in implementation of this data analysis.

B. ipynb

1) Data Preprocessing:

- Check the data set for any missing or null values. Figure 16 is the code snippet to perform this.
- Drop columns that are irrelevant to our analysis and model building, such as CustomerID, as 15 shows it is a random generated identification code, which is irrelevant to our project.
- Encode all the string variables. (No becomes 0 and Yes becomes 1 and so on...).

```
# Read the contents of the data
churn_data
```

	customerID	gender	SeniorCitizen
0	7590-VHVEG	Female	0
1	5575-GNVDE	Male	0
2	3668-QPYBK	Male	0
3	7795-CFOCW	Male	0
4	9237-HQITU	Female	0

Fig. 15: CustomerID.

This Figure 15 just shows the data set in python.

```
#the customerId is a random number+alphabets which is ir
churn_data = churn_data.drop('customerID', axis=1)

# check for any missing or null data in the dataset
churn_data.isnull().sum()

gender            0
```

Fig. 16: Code snippet of column drop and null value check.

This figure 16 shows the execution of dropping customerID column and checking for any null data in the dataset. CustomerID is a mix of 5 letters and 4 numbers.

2) Data Analysis: With the help of python library packages such as 'panda' and 'numpy', data loading, statistical analysis is done on the dataset. Data Visualization python library packages such as 'matplotlib.pyplot', 'seaborn library' are used for visual graphs and plots of the dataset. Later the data set is transformed to be fit for ML modelling using from sklearn.preprocessing import StandardScaler.

3) Performance Metrics: Define a function with all the performance metrics calculated, which can be called after every model. Figure 17 is the code snippet for performance metrics such as accuracy, f1 score, recall, and precision.

```
def performance_metrics(y_pred):
    print("Precision : ", round(precision_score(Y_test, y_pred, average = 'micro'),4))
    print("Recall : ", round(recall_score(Y_test, y_pred, average = 'micro'),4))
    print("Accuracy : ", round(accuracy_score(Y_test, y_pred),4))
    print("F1 Score : ", round(f1_score(Y_test, y_pred, average = 'micro'),4))
    cm = confusion_matrix(Y_test, y_pred)
    print("\n", cm)
    print("\n")
    print("****27 + "\n" + " "* 16 + "Classification Report\n" + "****27)
    print(classification_report(Y_test, y_pred))
    print("****27+"\n")

    cm = ConfusionMatrixDisplay(confusion_matrix = cm, display_labels=['Churn-NO', 'Churn-YES'])
    cm.plot()
```

Fig. 17: Code snippet of performance metrics function.

The Figure 17 shows the performance-metrics function that takes value of y prediction this executes the statistical analysis of the data like precision, recall and f1 score and classification report. These are basic units of statistical analysis.

4) Machine Learning Models: For this project, four machine learning models are selected. Encoded dataset and dimension reduced encoded dataset are used separately on the selected Machine learning models. After modelling the predicted values performance is assessed through the performance metrics function defined earlier. We generate confusion matrix for each model.

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive (TP)

Fig. 18: Confusion Matrix concept

The Figure 18 shows the correlation heat map for churn data, this includes all the columns in the dataset.

Model Training: The encoded dataset is split into training and testing dataset. Similarly Dimension reduced encoded dataset is split into training and testing dataset separately.

Logistic Regression: Logistic regression is a classification model that is used to classify the customer churn. Use sklearn.linear_model import LogisticRegression to build the model. Fit the training dataset into the model. Predict and run performance metrics function by calling predicted values into it.

GradientBoosting: Use from sklearn.ensemble import GradientBoostingClassifier to build the model. Fit the training dataset into the model. Predict and run performance metrics function by calling predicted values into it.

Random Forest: Use sklearn.ensemble import RandomForestClassifier. Fit the training dataset into the model. Predict and run performance metrics function by calling predicted values into it.

Neural Network: For the neural network be importance of flow and keras from tensorflow here we have opted for sequential in Keras which implies that output has only one value we have defined first input layer with 19 neurons, there are two hidden layers which have receiving neurons of 15 and 10. The activation function used in the input layer and hidden layers is ReLU and the output layer activation function is defined by sigmoid. ReLU is chosen because of it's computation of the backpropagation of neural networks and introducing non-linearity into the model. Sigmoid function transforms the value between 0 and 1 hence it has been chosen as an activation function for the outer layer. For the compilation of neural network we have used Adam Optimizer, binary cross entropy as loss, and accuracy as the performance metrics. We have trained the model for 60 epochs, we have chosen 60 epochs because it has been observed that the model on evaluating converges at epoch 56 for both kinds of dataset and the nearest round value is 60.

ROC-AUC Curve: The package sklearn.metrics has library roc_auc_score, which is imported into python code. This calculates the area under the receiver operating characteristics curve. Receiver operating characteristics a.k.a ROC is the graphical representation of the true positive rate to the false positive rate at different thresholds. This is also a part of performance metrics. Roc curve is generally plotted for binary classification models. The higher the area under the Curve the better is the model. The same package has roc_curve library which needs to be imported to record the false positive rate and true positive rate. This is inturn used to plot the ROC curve.

Precision-Recall Curve: Import precision_recall_curve, this calculates the Precision of the recall value at each probable threshold. These values are, in turn used to plot precision recall curve. Import auc from sklearn.metrics, to calculate the area under the precision-recall curve.

V. Preliminary Results (Testing and Deployment)

A. Dimension Reduction

The scree plot output obtained by Dimension reduction in IBM SPSS, gives the components plotted against their eigenvalues. The components with significant eigen value are chosen, as it implies that these features are the deciding components for the dependent variable. Based on the comparison of the scree plot with the correlation heatmap, figure 19 and figure 20, for a confidence interval of 95% [9], gives 9 components. This implies a dimensional reduction from 20 features to 9.

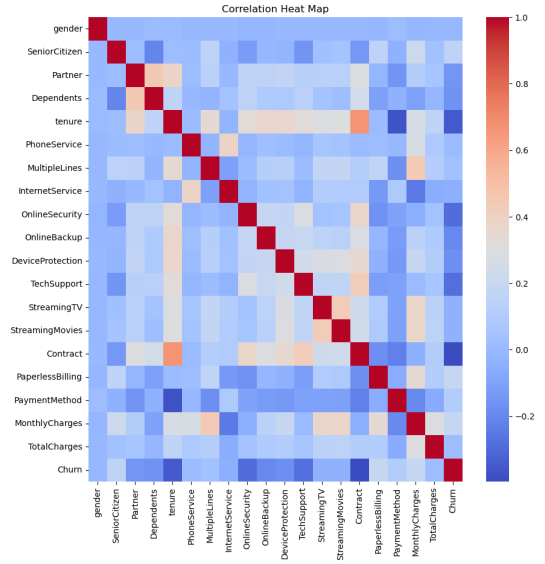

Fig. 19: Correlation heatmap for the churn dataset.

This Figure 20 is a scree plot, which is used to determine the optimal number of principal components to use for dimension reduction. It is plotted using eigenvalues and component numbers.

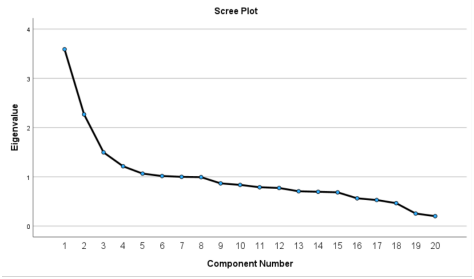

Fig. 20: Scree plot for dimension reduction done in IBM SPSS

The Figure 21 up until Figure 28 shows the confusion matrices of all machine learning algorithms tested on the dataset.

B. Logistic Regression

Logistic Regression

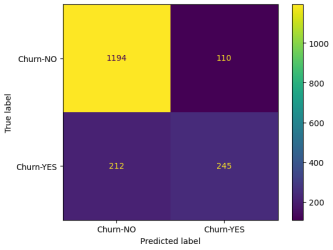

Fig. 21: Confusion matrix of Logistic Regression (without PCA).

From Figure 21, values for the confusion matrix for logistic regression are as follows:

True Negatives (TN): 1194
True Positives (TP): 245
False Positives (FP): 110
False Negatives (FN): 212
The accuracy of the model is 0.817, which means that 81.7% of the predictions were correct. The precision of the model is 0.69, which means that 69% of the predictions that were labeled as churn yes were actually churn yes. The recall of the model is 0.8171, which means that 81.71% of the customers who actually churned were correctly identified as churners.

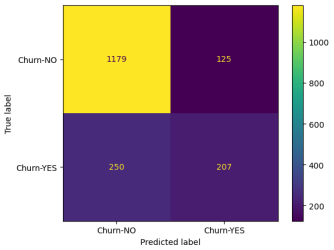

Fig. 22: Confusion matrix of Logistic Regression. (with PCA)

From Figure 22, the values for the confusion matrix for logistic regression with PCA are as follows:
True Negatives (TN): 1179
True Positives (TP): 207
False Positives (FP): 125
False Negatives (FN): 250
This shows that this model has a high precision of 83%, and a high recall of 90, which means the algorithm returns more relevant results.

C. Gradient Boosting

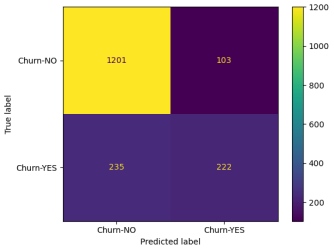

Fig. 23: Confusion matrix of Gradient Boosting.

From Figure 23, the values for the confusion matrix for Gradient Boosting are as follows:
True Negatives (TN): 1201
True Positives (TP): 222
False Positives (FP): 103
False Negatives (FN): 235
This model has high precision and accuracy of 84% and 80.81%.

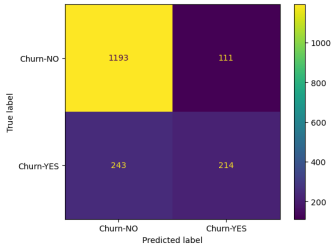

Fig. 24: Confusion matrix of Gradient Boosting. (with PCA)

From the Figure 24 the values for the confusion matrix for Gradient Boosting with PCA are as follows:
True Negatives (TN): 1193
True Positives (TP): 214
False Positives (FP): 111
False Negatives (FN): 243
This model has high precision and recall of 83% and 91%, also an accuracy of 79.9%.

D. Random Forest

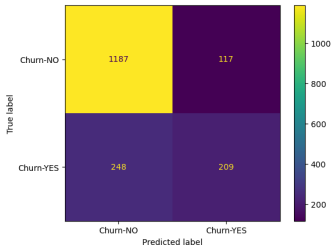

Fig. 25: Confusion matrix of Random Forest

From the Figure 25 the values for the confusion matrix for Random Forest are as follows:
True Negatives (TN): 1200
True Positives (TP): 208
False Positives (FP): 104
False Negatives (FN): 249
This model has precision and recall of 83% and 91%, also an accuracy of 80%.

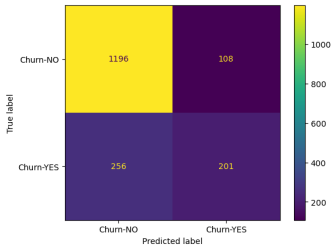

Fig. 26: Confusion matrix of Random Forest. (with PCA)

From the Figure 26 values for the confusion matrix for Random Forest with PCA are as follows:
True Negatives (TN): 1190
True Positives (TP): 199
False Positives (FP): 114

False Negatives (FN): 258
This model has a precision value of 82% (on the highest side), and recall value of 91% on the upper recall value.

E. Neural Network

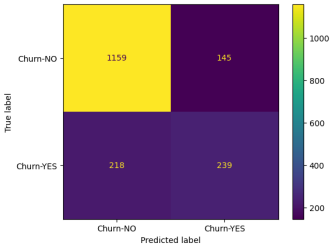

Fig. 27: Confusion matrix of Neural Network

From the Figure 26 values for the confusion matrix for Rneural network without PCA are as follows:
True Positives (TP): 239
True Negatives (TN): 1159
False Positives (FP): 145
False Negatives (FN): 218
This model has a precision value of 84% (on the highest side), and recall value of 89% on the upper recall value.

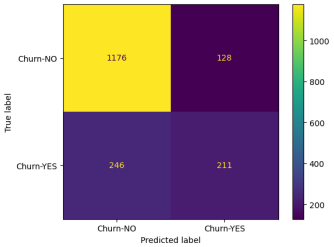

Fig. 28: Confusion matrix of Neural Network. (with PCA)

From the Figure 26 values for the confusion matrix for neural network with PCA are as follows:
True Positives (TP): 211.
True Negatives (TN): 1176
False Positives (FP): 128
False Negatives (FN): 246
This model has a precision value of 83% (on the highest side), and recall value of 90% on the upper recall value.

F. Accuracy of the models

TABLE I: Accuracy Values of the ML models

Model	Without PCA	With PCA
Logistic Regression	0.8171	0.7871
Gradient Boosting	0.8081	0.799
Random Forest	0.7933	0.788
Neural network	0.7939	0.7876

High Accuracy is observed from Logistic Regression (without PCA): 0.8171.
Lowest Accuracy is from Logistic Regression (with PCA): 0.7871.

G. Precision of the models

TABLE II: Precision Values of the ML models

Model	Without PCA			With PCA		
	0	1	Micro Avg	0	1	Micro Avg
Logistic Regression	0.85	0.69	0.8171	0.83	0.62	0.7871
Gradient Boosting	0.84	0.68	0.8081	0.83	0.66	0.799
Random Forest	0.83	0.64	0.7933	0.82	0.64	0.7888
Neural Network	0.84	0.62	0.73	0.86	0.57	0.73

Highest Precision is observed at Logistic Regression (without PCA):
Churn OFF(0) Precision: 0.85
Churn ON (1) Precision: 0.69
Micro Average: 0.8171
The Lowest Precision is observed at Logistic Regression (with PCA):
Churn OFF(0) Precision: 0.83
Churn ON (1) Precision: 0.62
Micro Average: 0.7871

H. f1 score of the models

TABLE III: f1 scores of the ML models

Model	Without PCA			With PCA		
	0	1	Micro Avg	0	1	Micro Avg
Logistic Regression	0.88	0.60	0.8171	0.86	0.52	0.7871
Gradient Boosting	0.88	0.57	0.8081	0.87	0.55	0.7978
Random Forest	0.87	0.53	0.70.8018	0.87	0.57	0.7933
Neural Networks	0.86	0.57	0.7939	0.86	0.53	0.7876

Highest f1-score is observed at Logistic Regression (without PCA):
Churn OFF(0) f1 score: 0.88
Churn ON (1) f1 score: 0.60
Micro Average: 0.8171
The Lowest f1-score is observed at Logistic Regression (with PCA):
Churn OFF(0) f1 score: 0.86
Churn ON (1) f1 score: 0.52
Micro Average: 0.7871

I. Recall Score of the models

TABLE IV: Recall scores of the ML models

Model	Without PCA			With PCA		
	0	1	Micro Avg	0	1	Micro Avg
Logistic Regression	0.92	0.54	0.8171	0.90	0.45	0.7871
Gradient Boosting	0.92	0.49	0.8081	0.91	0.47	0.7978
Random Forest	0.92	0.47	0.8018	0.91	0.45	0.7933
Neural Networks	0.89	0.52	0.7939	0.90	0.46	0.7876

Highest recall score is observed at Logistic Regression (without PCA):
Churn OFF(0) recall score: 0.92
Churn ON (1) recall score: 0.54
Micro Average: 0.8171
The Lowest recall score is observed at Neural Networks (with PCA):
Churn OFF(0) recall score: 0.90
Churn ON (1) recall score: 0.46
Micro Average: 0.7876

VI. ANALYSIS

A. ROC-AUC Curve

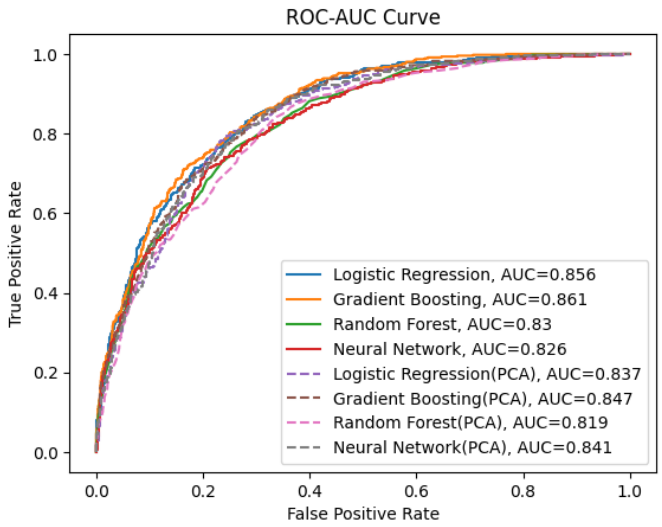

Fig. 29: ROC-AUC Curve

ROC curve gives the graphical representation of the performance metric of the binary classification. TPR vs FPR is plotted for different classification thresholds. Higher area implies better model provided the data is not imbalanced.

B. Recall Analysis

Precision: It is used to measure the percentage of correctly identified events out of all the predicted values. Where as recall attempts to calculate the actual number of positives that were correctly identified. Recall is also known as sensitivity. Precision and recall have an inverse relation with each other. When data is heavily imbalanced, precision-recall curve is considered as deciding performance metric compared to all other including ROC [11].

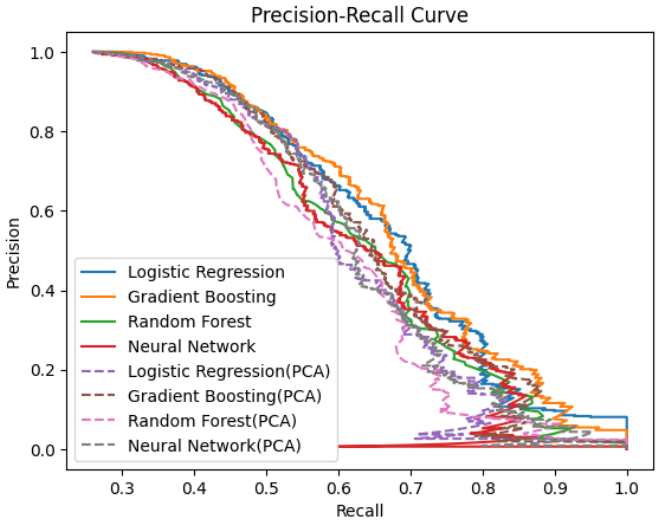

Fig. 30: Precision-Recall Curve

TABLE V: Area under the curve(AUC) values for Receiver Operating Characteristics(ROC) curve and Precision-Recall Curve

ML-Model	ROC-AUC	Precision-Recall AUC
Logistic Regression	0.856	0.676
Gradient Boosting	0.861	0.684
Random Forest	0.83	0.636
Neural Network	0.826	0.628
Logistic Regression (with PCA dataset)	0.837	0.622
Gradient Boosting (with PCA dataset)	0.847	0.65
Random Forest (with PCA dataset)	0.819	0.603
Neural Network (with PCA dataset)	0.841	0.633

Table V displays area values under the ROC and precision-recall curve for all 8 machine learning models. It is desirable to have the largest area as that implies that the model has a higher true positive rate compares to others. Though ROC is a good performance metrics, as our data is imbalanced, meaning unequal YES and NO. It has bias introduced, thus precision-recall curve takes precedent in determining better model. Data from V implies that gradient boosting classifier, for undiminished dataset features, is a better predictor.

VII. CONCLUSION

Based on the performance parameters, for the undimension reduced dataset Logistic regression(LR) has better accuracy, and LR has better precision to predict 'NO' and 'YES'

churn. LR, GB, RF gave equal recall or sensitivity scores to predict 'NO'. But only LR has a higher score for predicting 'YES.' Thus, it has been concluded that Gradient Boosting without principal component Analysis has a better prediction rate compared to other models. Though doing PCA reduces the features complexity, it is found that the overall accuracy decreases. Our criteria is to select best performing ML model, though if the features are to be compromised, the accuracy decreases by 0.0091 accuracy score or 0.91%.

a) Issues and concerns:: While data prediction helps us to predict data it only gives a probabilistic value rather than absolute. The dataset for our project has only 20 features but there are churn datasets with greater dimensionality. When it comes to that it is better to compromise the accuracy by a few degrees and perform dimension reduction.

VIII. Project Management

Current Implementation Status: Complete

A. Work Completed

The considered project is deemed to have achieved the proposed goal. The code and document have been uploaded to GitHub link https://github.com/KShinyShamma/Machine-Learning

- Literature review and Data preprocessing - by Hanu Vamsi Putta (11704879), read papers for related work, and did data encoding for data preprocessing.
- Descriptive Statistics and Data Visualization - by Sai Bhavani Shankar Chintapalli (11724519). Plotted different graphical representations of data.
- Principal Component Analysis(PCA) Using SPSS - by Sreenivas Varma Chamarthi (11725404) Used IBM SPSS in order to reduce the total features of the data set.
- Build the following Machine learning models - by Shiny Shamma Kota(11698716). Leveraged sklearn documentation to perform these machine learning models.
    1) Logistic Regression - for both of the datasets.
    2) Gradient Boosting - for both of the datasets.
    3) Random Forest - for both of the datasets.
    4) Neural Network(50%) - For PCA performed dataset
- Build the Neural Network model without PCA(50%) - by Sreenivas Varma Chamarthi (11725404)
- ROC-AUC curve - by Hanu Vamsi Putta (11704879). Using sklearn performed ROC-AUC in colab.
- Recall Analysis in performance metrics - by Sai Bhavani Shankar Chintapalli (11724519). Performed Precision - Recall Graph in google colab.

References

[1] S. L. Cabrera-Luján et al., "Impact of corporate social responsibility, business ethics and corporate reputation on the retention of users of third-sector institutions," Sustainability, vol. 15, no. 3, p. 1781, 2023. doi:10.3390/su15031781

[2] J. Frankenfield, "Churn rate: What it means, examples, and calculations," Investopedia, https://www.investopedia.com/terms/c/churnrate.asp (accessed Nov. 19, 2023).

[3] A. Raj and D. Vetrithangam, "Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 139-144, doi: 10.1109/CISES58720.2023.10183530.

[4] ZY Wang, "Design and Implementation of Telecom Customer Loss Prediction Model Based on Particle Swarm Optimization Algorithm", Beijing: University of Chinese Academy of Sciences (School of Artificial Intelligence Chinese Academy of Sciences), 2019.

[5] X. Hu, Y. Yang, L. Chen and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 129-132, doi: 10.1109/ICCCBDA49378.2020.9095611.

[6] M. Saghir, Z. Bibi, S. Bashir and F. H. Khan, "Churn Prediction using Neural Network based Individual and Ensemble Models," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2019, pp. 634-639, doi: 10.1109/IBCAST.2019.8667113.

[7] L. Lin, J. Ma, X. Ye and X. Xu, "Mechanical fault prediction based on principal component analysis," The 2010 IEEE International Conference on Information and Automation, Harbin, China, 2010, pp. 2258-2262, doi: 10.1109/ICINFA.2010.5512443.

[8] BlastChar,"Telco customer churn,"Kaggle, https://www.kaggle.com/datasets/blastchar/telco-customer-churn (accessed Nov. 19, 2023).

[9] "Principal component analysis(pca),"GeeksforGeeks, https://www.geeksforgeeks.org/principal-component-analysis-pca/ (accessed Nov. 19, 2023).

[10] "Guide to AUC ROC curve in machine learning," GeeksforGeeks, https://www.geeksforgeeks.org/auc-roc-curve/ (accessed Dec. 3, 2023).

[11] J. Brownlee, "ROC curves and precision-recall curves for imbalanced classification," MachineLearningMastery.com, https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/ (accessed Dec. 3, 2023).