Telecom Customer Churn Prediction, COMP7810

Group 33

I. Introduction

It has been estimated that churn rate in telecom industry is around 30% [1]. The acquisition of new customers is 5-10 times costlier than retaining the existing ones [2]. In the telecom industry, people have multiple options and the risk of churn is especially high. So, this show the importance of reducing the churn rate in telecom industry. Predicting customer churn with machine learning models is a potential solution to this problem. In this project, multiple machine learning models were built to predict churn from relevant customer quantitative and qualitative variables. The results on the open source dataset show the possibility of achieving this goal. Support vector machine (SVM) model demonstrated the best performance with 85.73 % F-1, 89.76 % precision, 82.04 % recall, 85.98 % AUC, and 85.84% accuracy. The top 3 important features were total charges, tenure, and contract type. Analysis of these features reveal that customers are likely to leave at an earlier stage of tenure. Also, people who are signed up for monthly subscription have a higher chance of leaving the service. Furthermore, the data show that majority of customers who are more likely to leave aren't profitable ones, showing that problem is less severe than it appears. The rest of the report is structured as follows. The problem definition is defined in Section 2. Section 3 contains methodology that describes exploratory data analysis, preprocessing, model selection, and feature analysis. Section 4, includes the findings and discussion. Finally, Section 5 consist of conclusion.

II. Problem Statement Definition

The main problems of this project can be stated in the following way:

- **Evaluate the effectiveness of machine learning algorithms for predicting the churn from customer data:** The problem in this project is to build a model to predict churn in telecom from relevant features.

- **Identify the most important features for predicting churn:** It is not only important to predict but also identify top relevant features to reduce the churn rate.

- **Discovery of business insights for decreasing churn rate by analyzing important features**

III. Methodology

A. Dataset Exploration Analysis

The dataset contains 7043 customer records with 21 features[1]. Last feature indicates whether customer left or not. The rest of the features describe relevant information such as gender and types of service the customer has subscribed for. Tables 1 and 2, show the descriptive statistics of 3 quantitative and 16 qualitative variables in the dataset. Figure 1 displays the box plot for quantitative variables and indicates that there is no outlier in quantitative variables. Qualitative variables don't have extreme values as shown on the Table 2. The dataset contains 7010 rows after removing duplicates and rows with missing features.

[1]https://www.kaggle.com/datasets/blastchar/telco-customer-churn

TABLE I: Descriptive Statistics of Quantitative Features

Feature	Range	Mean (SD)
Tenure (months)	1 - 72	32.52 (24.52)
Monthly Charges	18.25 - 118.75	64.89 (30.06)
Total Charges	18.80 - 8684.80	2290.35 (2266.82)

TABLE II: Descriptive Statistics of Qualitative Features

Features	Values	Frequencies
Gender	Male, Female	3535, 3475
Senior Citizen	Yes, No	1141, 5869
Partner	Yes, No	3393, 3617
Dependents	Yes, No	2099, 4911
Phone Service	Yes, No	6330, 680
Multiple Lines	Yes, No	2967, 3363
Internet Service	Fiber optic, DSL, No	3090, 2414, 1506
Online Security	Yes, No, No internet service	2015, 3489, 1506
Online Backup	Yes, No, No internet service	2425, 3079, 1506
Device Protection	Yes, No, No internet service	2418, 3086, 1506
Tech Support	Yes, No, No internet service	2040, 3464, 1506
Streaming TV	Yes, No, No internet service	2703, 2801, 1506
Streaming Movies	Yes, No, No internet service	2731, 2773, 1506
Contract	Month-to-month, One year, Two year	3853, 1472, 1685
Paperless Billing	Yes, No	4158,2852
Payment Method	Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)	2359, 1588,1542,1521

B. Preprocessing

Multiple pre-processing steps were performed before the model selection. First, 11 rows with missing values are removed as the number of such rows are very low and there is no need for using imputation techniques. Also, 22 duplicate rows were omitted. Customer ID column is excluded, as it is not useful for modeling. All categorical features were one-hot encoded. Besides, there exist a class imbalance problem as the number of churned customers is 1857, while the number of loyal customers is 5153. The SMOTE technique was used to address this problem. SMOTE technique address the problem by generating synthetic data for minority class [3]. Furthermore, features have very different range as shown in Table 1 and all features were normalized to have 0 mean and unit variance. The dataset was split in 80/20 ratio for training and testing. All the features were included to build to machine learning model as the dataset is large enough. Pre-processing steps were carried out by Python.

C. Model Selection

Machine learning models were trained to predict the churn from customer data. The dataset size and type influenced the model selection process. The dataset size is relatively large and complex models can be

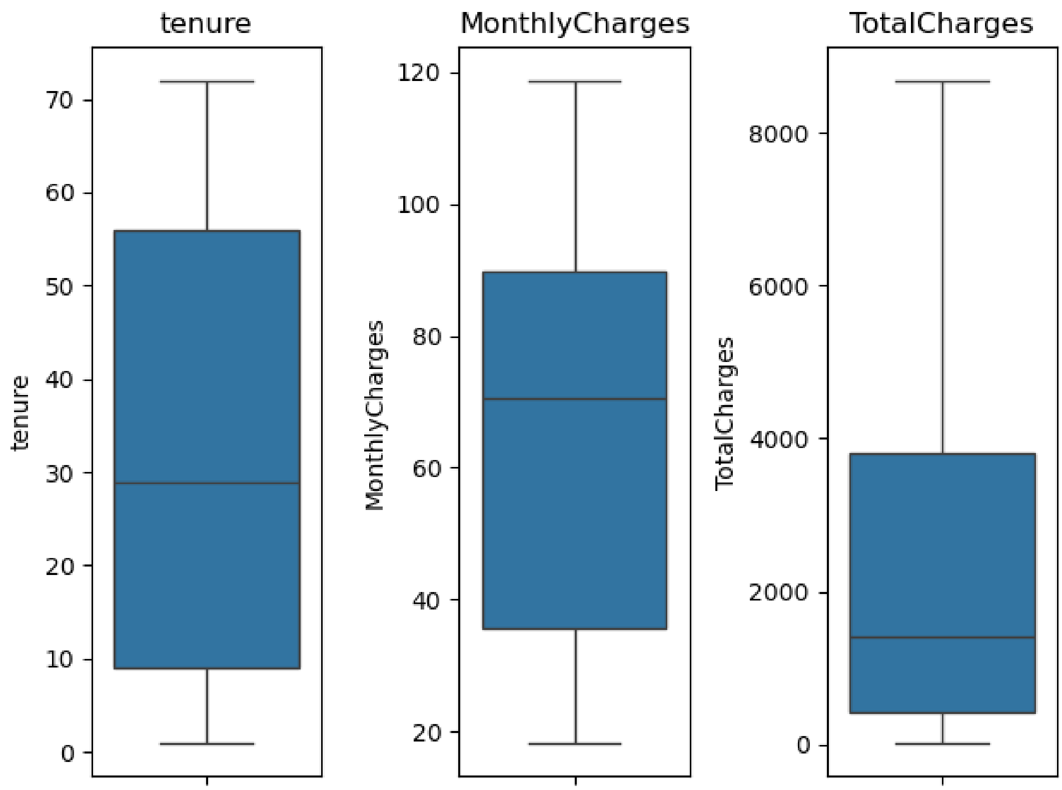


Fig. 1: Box plot for quantitative features

TABLE III: Comparison of ML models

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F-1 (%)	AUC (%)
SVM	**85.84**	**89.76**	82.04	**85.73**	**85.98**
Random Forest	85.26	87.61	**83.35**	85.43	85.33
Decision Tree	80.70	81.74	80.82	81.28	80.69

selected. Since the dataset contains many categorical features, tree-based models had a preference. Also, selection was based on the effectiveness of models on previous studies [4] [5]. In this project, SVM, Random Forest, and Decision Tree algorithms were selected. Python and sklearn package were used to build the models.

IV. Findings and Discussion

A. Model performance

All models have good performance in terms of important metrics showing the possibility of predicting churn from customer data. The dataset is imbalanced and accuracy isn't enough to assess the performance of a model. Models were also evaluated in terms of precision, recall, F-1, and AUC score to identify the most suitable model for churn detection. Table 3, contains the model performance, and the confusion matrix for each model is shown in Figure 6. The SVM model has the highest performance in terms of quantitative metrics achieving 85.73 % F-1, 89.76 % Precision, 82.04 % Recall, 85.98 % AUC, and 85.84% Accuracy. Figure 2 provides the visual comparison of models.

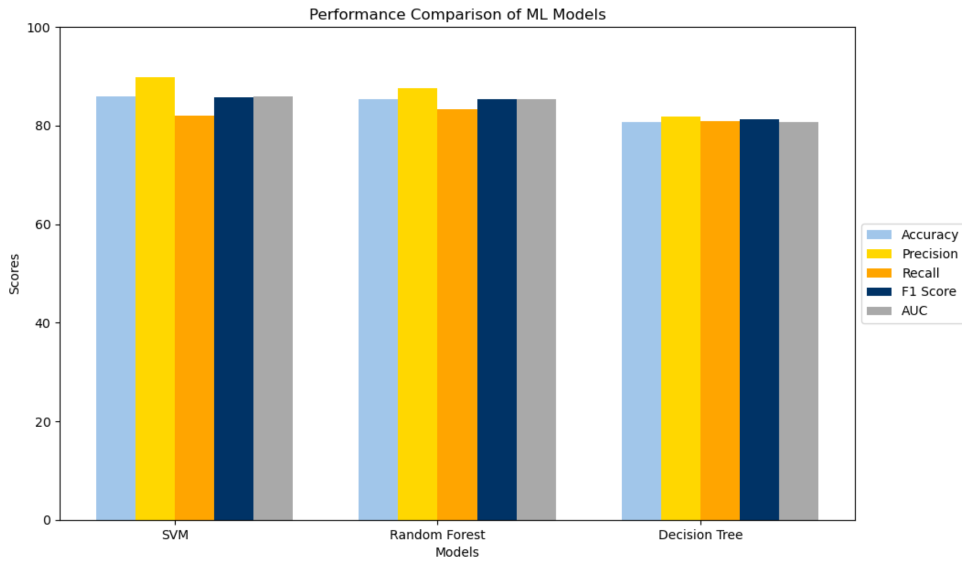


Fig. 2: Performance Comparison of ML Models

B. Feature importance analysis

The Table 4 contains the importance of features derived from Random Forest Algorithm. The importance of features calculated as a gain in information where impurity is defined as gini index [6]. Top 3 features are total charges, tenure (months) and the whether contract is month to month based. Tenure refers to total number of months the customer stayed with the company. Figures 3,4,5 show the distribution of top 3 features in terms of churn variable. The total charges plot show that customers are likely to quit if their total charges are low. This shows that the majority of churned customers are not important customers. The tenure plot shows that the customers are more likely to churn during the first few months. So this shows the importance of keeping the customers at the beginning. Besides, the distribution of binary variable contract (month-to-month) show that customers tend to leave if they sign up for monthly contract. So, companies can try to sign up customers for one year or two year contracts to decrease the churn rate.

V. Conclusion

Retaining the customers who are likely to leave is crucial for the telecom industry companies. Machine learning algorithms are possible solution to this issue. In this project, multiple machine learning models were developed to assess their effectiveness. The results show the possibility of machine learning models to predict the churn. Also, multiple business insights were derived by analyzing the important features in the model.

References

[1] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of customer churn behavior in the telecommunication industry using machine learning models," *Algorithms*, vol. 17, no. 6, p. 231, 2024.

[2] Y. Yulianti and A. Saifudin, "Sequential feature selection in customer churn prediction based on naive bayes," in *IOP conference series: materials science and engineering*, vol. 879, no. 1. IOP Publishing, 2020, p. 012090.

[3] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting covid-19 mortality," *BMC medical informatics and decision making*, vol. 22, no. 1, p. 2, 2022.

[4] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022.

[5] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[6] X. Yuan, S. Liu, W. Feng, and G. Dauphin, "Feature importance ranking of random forest-based end-to-end learning algorithm," *Remote Sensing*, vol. 15, no. 21, p. 5203, 2023.

TABLE IV: Importance of features

Row	Feature	Importance	Row	Feature	Importance
1	TotalCharges	0.1107	24	StreamingMovies_Yes	0.0117
2	Tenure	0.1010	25	DeviceProtection_Yes	0.0112
3	Contract_Month-to-month	0.0880	26	Dependents_Yes	0.0108
4	MonthlyCharges	0.0875	27	TechSupport_Yes	0.0107
5	OnlineSecurity_No	0.0668	28	PaymentMethod_Credit card (automatic)	0.0105
6	PaymentMethod_Electronic check	0.0569	29	Contract_One year	0.0104
7	TechSupport_No	0.0540	30	StreamingTV_Yes	0.0102
8	OnlineBackup_No	0.0305	31	MultipleLines_No	0.0101
9	DeviceProtection_No	0.0262	32	SeniorCitizen_No	0.0100
10	InternetService_Fiber optic	0.0260	33	PaymentMethod_Mailed check	0.0098
11	PaperlessBilling_Yes	0.0253	34	StreamingMovies_No	0.0093
12	Partner_No	0.0180	35	StreamingTV_No	0.0093
13	Dependents_No	0.0164	36	InternetService_DSL	0.0073
14	Gender_Female	0.0158	37	MultipleLines_No phone service	0.0032
15	SeniorCitizen_Yes	0.0150	38	PhoneService_No	0.0032
16	Gender_Male	0.0149	39	PhoneService_Yes	0.0026
17	Partner_Yes	0.0147	40	OnlineBackup_No internet service	0.0026
18	Contract_Two year	0.0143	41	StreamingTV_No internet service	0.0023
19	PaperlessBilling_No	0.0128	42	TechSupport_No internet service	0.0022
20	OnlineBackup_Yes	0.0128	43	InternetService_No	0.0022
21	MultipleLines_Yes	0.0125	44	StreamingMovies_No internet service	0.0021
22	PaymentMethod_Bank transfer (automatic)	0.0120	45	DeviceProtection_No internet service	0.0021
23	OnlineSecurity_Yes	0.0118	46	OnlineSecurity_No internet service	0.0021

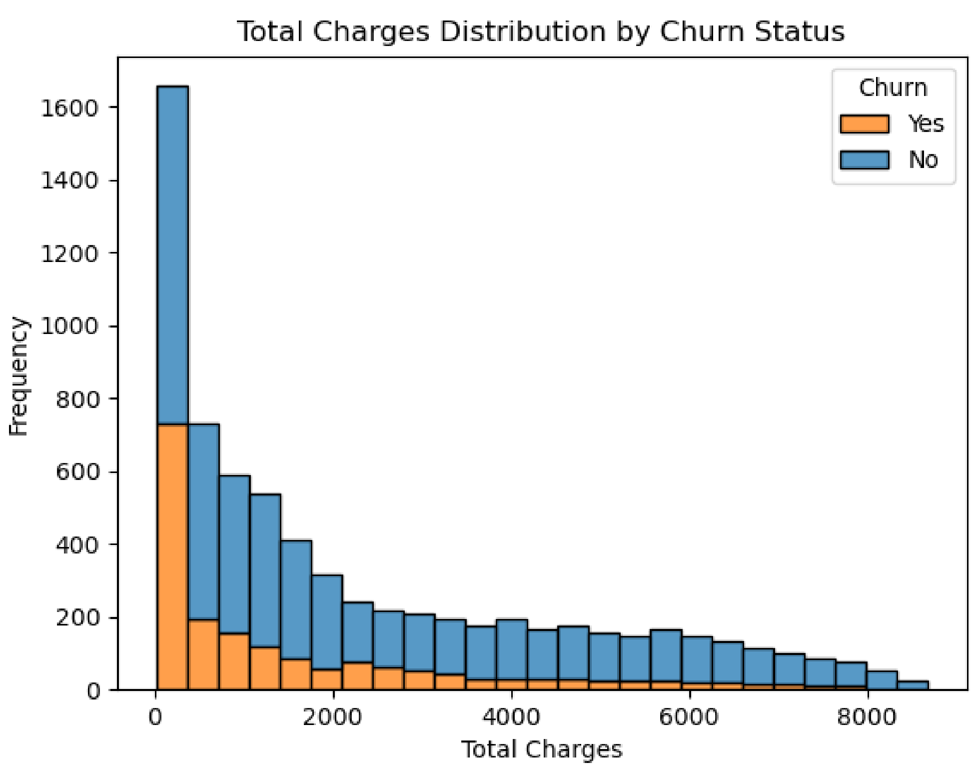


Fig. 3: Distribution of Total Charges

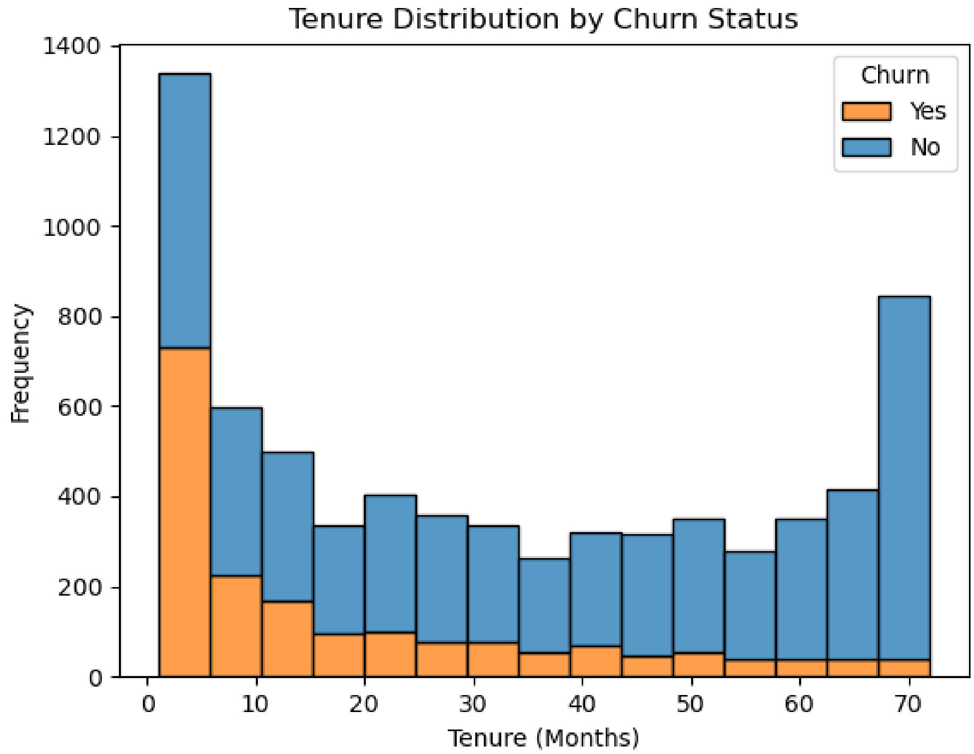


Fig. 4: Distribution of Tenure (Months)

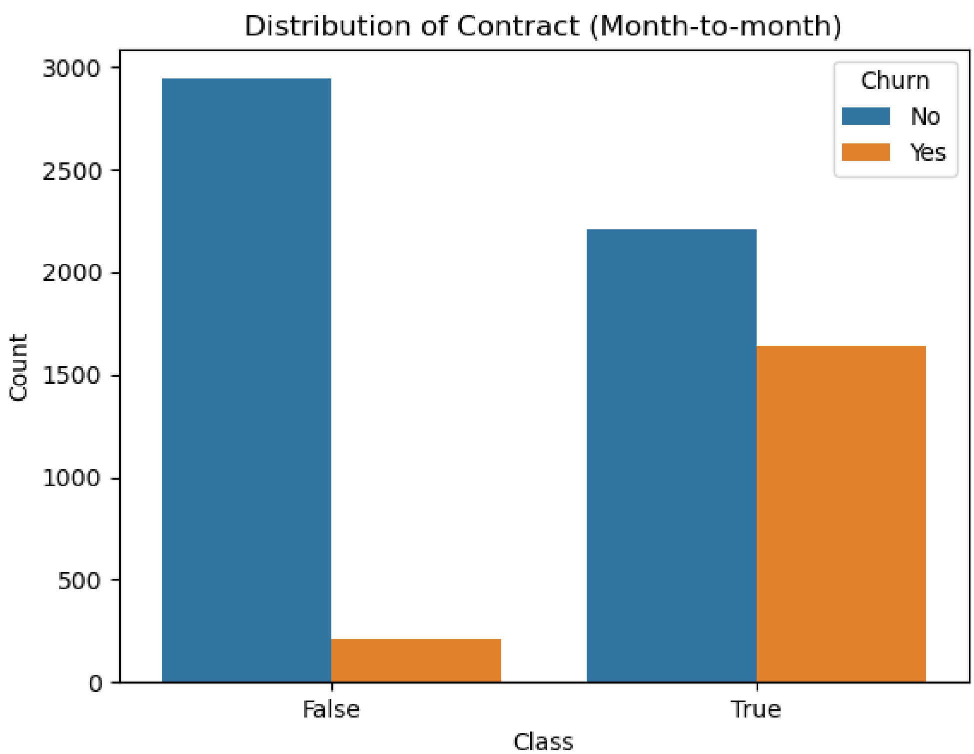
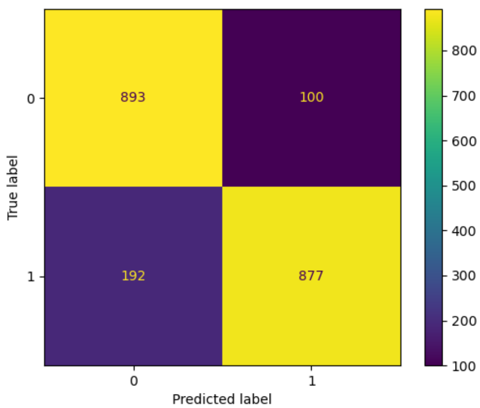
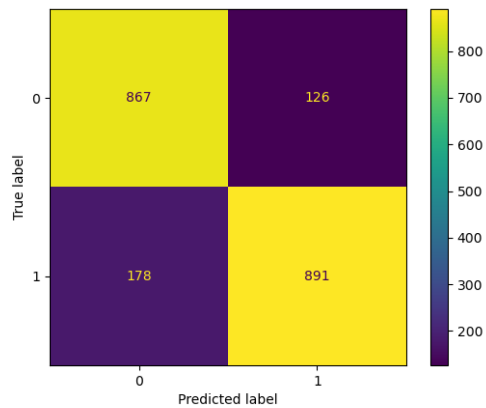


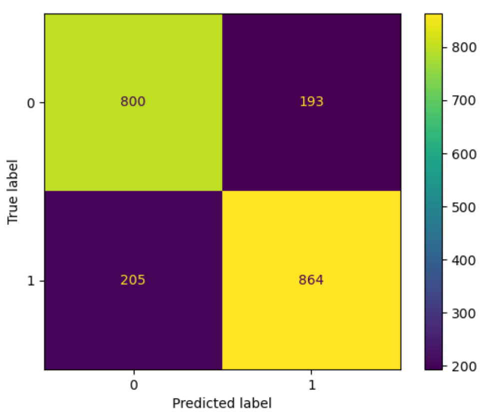
Fig. 5: Distribution of Contract (Month to month)



(a) SVM



(b) Random Forest



(c) Decision Tree

Fig. 6: Confusion matrix for ML models