School of Business and Management A350A0250

A220A0752 Analytics for Business

Professor, Azzurra Morreale

# Iranian Churn Prediction

Author

Prashant Shrestha

# Table of contents

# 1. Introduction

In Iran there are approximately five main telecom companies serving to consumers. Currently, the country is still transitioning from 4G to 5G technology. With a population of about 87 million people (2023, The World Bank), it is important for telecommunication companies to understand why their consumers would churn and switch over to a competitor (Jafari-Marandi, Denton, Idris, Smith, Keramati, 2020).

To understand the churn patterns in Iran's telecom industry, it is crucial to develop a prediction model that can uncover hidden patterns and help to predict the customers which will churn in future. (Jafari-Marandi et al., 2020). This will help the company to take effective actions to prevent customers from churning. The Iranian Churn Dataset comprises 12 months of telecom user's data recorded anonymously. The dataset features 13 attributes and 3150 observations. The 13 features are given below.

Anonymous Customer ID which records 3150 entries.

Call Failures: number of call failures.

Complaints: binary attribute, 0 for no complaints, 1 for complaint.

Subscription length: recorded in months.

Charge Amount: ordinal attribute from 0 to 9. 0 for the lowest amount and 9 for highest amount.

Seconds of use: total number of seconds a user spent on calls.

Frequency of Use: total number of calls.

Frequency of SMS: total number of text messages used.

Distinct called numbers: total number of distinct phone calls made.

Age Group: Ordinal attribute from 1 to 5. 1 for a younger age group and 5 for the oldest age group.

Tariff Plan: binary attribute. 1 for "pay as you go" and 2 for a contractual service.

Status: binary attribute. Refers to 1 for an active customer, 2 for non-active.

Churn: binary attribute. 1 churn, 0 non-churn. Class label

Customer Value: calculated value of a customer

## i. Problem definition and Objective

In this project, the researcher will focus on two major research questions "How effective classification models in predicting the churn for the Iranian dataset are?" and "What are the significant factors that contribute to churn in the company?'.

In this project, the researcher will primarily create and select a robust and effective classification model by comparing Logistic regression, Lasso Logistic Regression, Support Vector Machines and Classification Tree. The researcher will present how effectiveness of the classification models vary when predicting churn for the given data. The researcher will use various evaluation metrics like confusion matrix, accuracy, F1score, AUC and ROC curve to select the best classification model. Moreover, the researcher will perform exploratory and clustering analysis for understanding the characteristics of segmented customers and evaluate the clusters that are more likely to churn. This will help to understand how attributes influence churn rates and what are the drivers that consumers consider when making the decision to switch providers.

### ii. Research

In previous research relating to this dataset, Jafari-Marandi et al. (2020) proposed an analysis based on artificial neural networks which added a new dimension "Misclassification" for predicting churn. Misclassification was used to measure how successful a method was to find the best decisions for minimizing profit loss. It was proven that there is a 95% accuracy in results when predicting churn using Decision Tree, Artificial Neural Networks, and K-Nearest Neighbors methods (Keramati, Jafari-Marandi, Aliannejadi, Ahmadian, Mozaffari, Abbasi, 2014).

On the contrary, our project is limited to selection of best classification model for the dataset and exploration of the dataset using clustering and data analysis. The selection of the best classification model will be done through comparison of performance metrics like AUC, ROC, accuracy, precision, recall and F1-score. Moreover, the researcher aims to explain how a telecom company can use the project findings and the selected predictive model to increase customer retention and revenue generation.

## 2. Data Understanding and Exploration

In this section, a comprehensive analysis was conducted to explore and understand the data. This involved univariate analysis, multivariate analysis, and clustering analysis. These analytical techniques provide valuable insights in answering the research questions and contribution to the creation of a more effective model in subsequent sections. Before analysis, missing values are addressed by removing rows that contain missing values in any variable.

### i. Univariate Analysis

Table 1 and Figure 1 help to understand the distribution of the continuous variables. The dataset is not normally distributed. All the continuous variables except 'Subscription Length' are positively

skewed. This suggests that the data points in these variables are concentrated in the lower values of the distribution. However, the datapoints in the 'Subscription Length' are concentrated in towards the higher values of the distribution. Moreover, the box plots for each variable also depict the outliers present in the distribution. Therefore, the distribution must be normalized or standardized before the creation of the model.

*Table 1 Descriptive statistics for continuous variables*

| Variables | Min | Median | Mean | Max | Std. Dev. | Skewness |
|---|---|---|---|---|---|---|
| Call Failure | 0 | 6 | 7.6279 | 36 | 7.2639 | 1.0892 |
| Subscription Length | 3 | 35 | 32.5419 | 47 | 8.5735 | -1.2994 |
| Charge Amount | 0 | 0 | 0.9429 | 10 | 1.5211 | 2.5836 |
| Seconds of Use | 0 | 2990 | 4.4725 | 17090 | 7.1979 | 1.3213 |
| Frequency of Use | 0 | 54 | 69.4606 | 255 | 57.4133 | 1.1436 |
| Frequency of SMS | 0 | 21 | 73.1749 | 522 | 112.2376 | 1.9732 |
| Distinct Called Numbers | 0 | 21 | 23.5098 | 97 | 17.2173 | 1.0289 |
| Customer Value | 0 | 120.6750 | 170.9503 | 817.6500 | 172.6268 | 1.6616 |

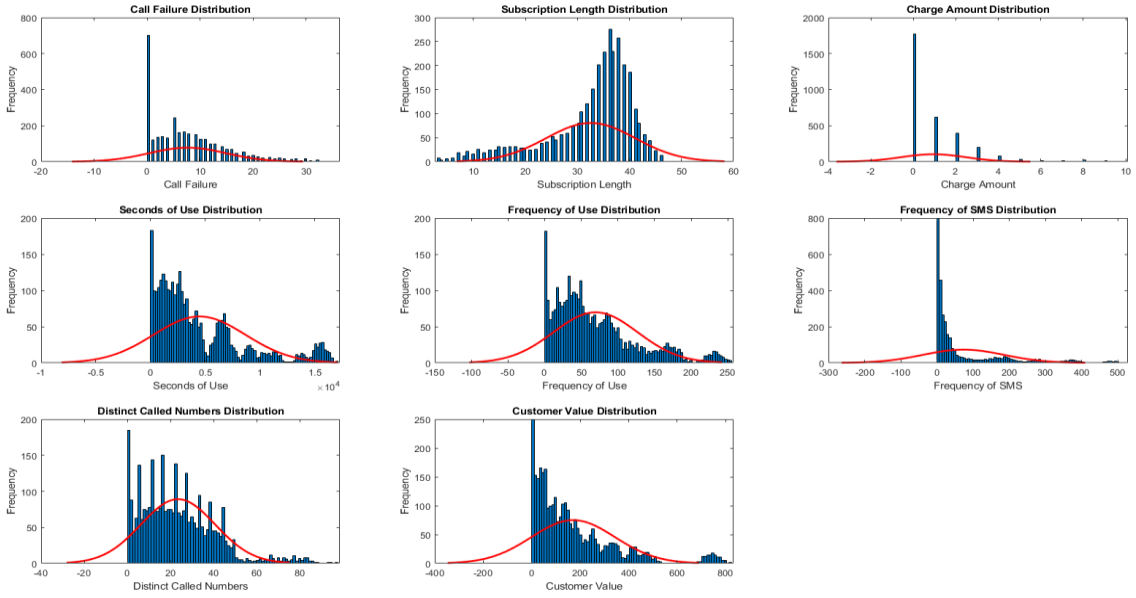


*Figure 1. Standard deviation, Skewness, and Kurtosis for all continuous variables.*

Figure 2 presents how categorical variables are distributed. Generally, there are more users who have no complaint. Moreover, most users prefer "pay-as-you-go" tariff plan. Furthermore, most users have not churned in the dataset. Similarly, there are more active users than non-active. We can also see the distribution of the Age Group variable classified from Young to Old. More users fall in the 2 and 3 categories representing comparatively young population.

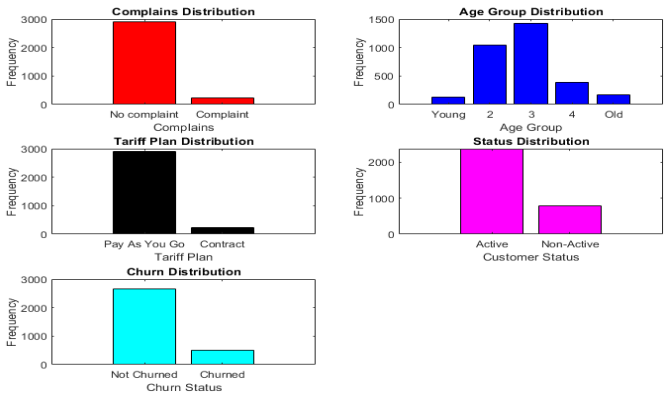


*Figure 2 Distribution of categorical variables*

## ii. Multivariate Analysis

In figure3, It is observed that 'Frequency of Use' and 'Seconds of Use' are highly positive correlated with a coefficient of 0.9465. Similarly, 'Customer Value' and 'Seconds of Use' are highly positive correlated with a coefficient of 0.8372. 'Distinct Called Numbers' and 'Seconds of Use' are also positive correlated with a coefficient of 0. 6765.Contrarily, 'Subscription Length' seems to have almost no correlation with all other variables with values close to 0.

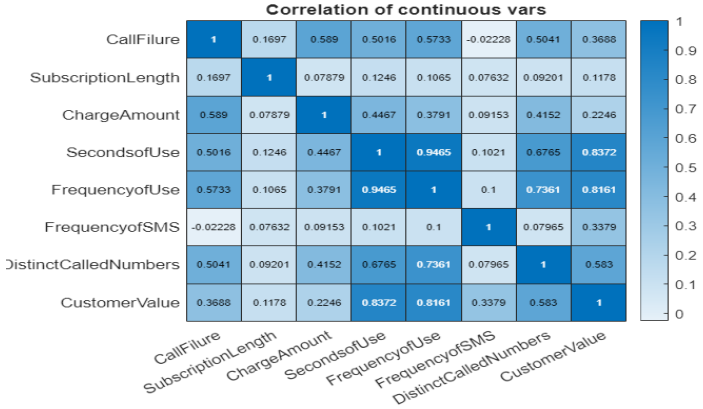


*Figure 3 Correlation of continuous variables.*

## iii. Clustering Analysis

This section presents the results derived from the clustering analysis using the K-means method. The clusters in the dataset were created using only the continuous variables. The variables were normalized using the z score standardization method. Figure 4 illustrates that the silhouette value is comparatively higher when there are 2 clusters so, the optimal number of clusters based is 2. Furthermore, Figure 4 reveals that cluster 2 has less customers who churned, whereas cluster 1 has a significantly higher number of customers who churned.

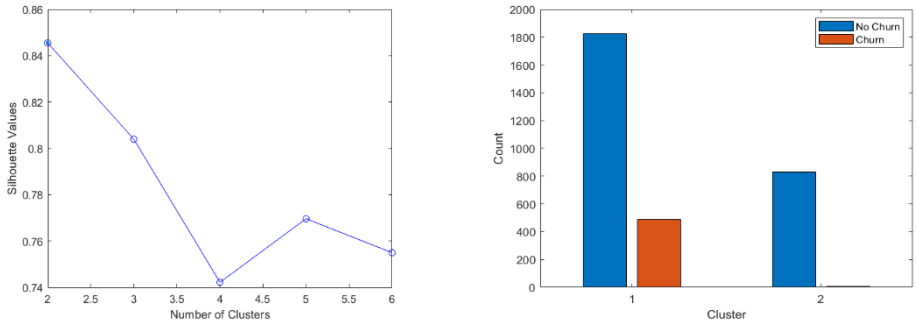


*Figure 4 Silhouette values and Bar graph of churn for each cluster*

Table 2 depicts that cluster 2 exhibits a higher number of call failures compared to cluster 1. Additionally, cluster 2 has a lower number of complaints than cluster 1. Moreover, cluster 2 demonstrates a higher average charge amount compared to cluster 2. Furthermore, users in cluster 2 tend to use the service more frequently for SMS and distinct calls.

Both clusters share similarities in terms of subscription length and age group. However, the mean and standard deviation of churn suggests that customers in cluster 1 are more likely to churn than those of cluster 2. Furthermore, customers in cluster 1 depict higher levels of inactivity compared to cluster 1 and they are less likely to use service for SMS and distinct calls. Similarly, the customers in cluster 2 are more likely to have contractual traffic plan than “pay as you go” traffic plan. Lastly, the customer value and charge amount of cluster 1 is comparatively lower than that of cluster 2.

*Table 2 Mean and SD of each cluster.*

| **Variables** | **Cluster 1** | | **Cluster 2** | |
|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** |
| Call Failure | 5.51 | 5.60 | 13.38 | 8.10 |
| Complaint | 0.09 | 0.29 | 0.03 | 0.17 |
| Subscription Length | 32.07 | 8.61 | 33.83 | 8.35 |
| Charge Amount | 0.50 | 0.87 | 2.14 | 2.14 |
| Seconds of Use | 2411.23 | 1769.13 | 10067.92 | 3757.83 |
| Frequency of Use | 41.64 | 27.80 | 144.99 | 48.38 |
| Frequency of SMS | 60.57 | 108.78 | 107.40 | 114.38 |
| Distinct Called Numbers | 16.81 | 13.04 | 41.70 | 13.67 |
| Age Group | 2.83 | 0.85 | 2.82 | 0.99 |
| Traffic Plan | 1.03 | 0.18 | 1.20 | 0.40 |
| Status | 1.34 | 0.47 | 1 | 0.00 |
| Churn | 0.21 | 0.41 | 0.01 | 0.10 |
| Customer Value | 98.08 | 79.15 | 368.76 | 200.41 |

## 3. Model Development

This section presents the process of developing classification models. It includes preprocessing of data, hyperparameter optimization and model evaluation. By following these steps, the section provides a comprehensive overview of the development process for classification models on the churn dataset, ensuring the selection of robust and effective models for churn prediction. Moreover, the reason behind the model choices made for the churn dataset in this project will also be discussed, considering factors such as interpretability and performance of the churn prediction task. The process of model development discussed in this section is depicted in figure 5.

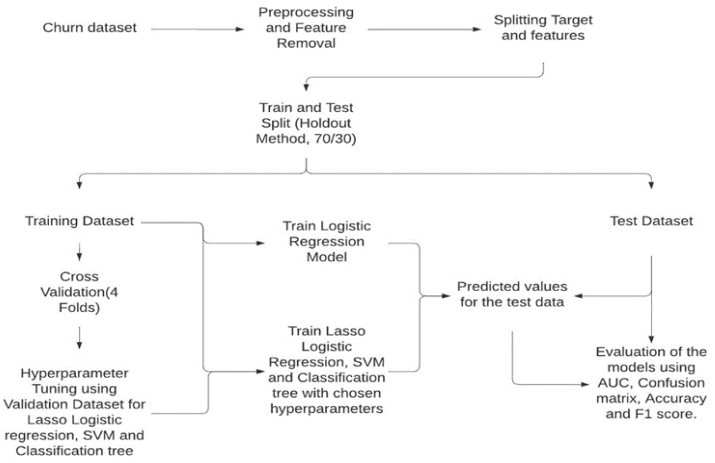


*Figure 5 Model development*

### i. Pre-processing data:

The process of pre-processing the dataset involves several steps to ensure data quality and prepare it for modeling. Initially, missing values are addressed by removing rows that contain missing values in any variable. Next, the target variable "Churn" and features such as "Complains", "Age Group", "Tariff Plan", and "Status" are converted into categorical variables to facilitate modeling.

Moreover, the dataset is divided into a target vector and a features vector. To improve the performance of the models and eliminate irrelevant information, the feature "Anonymous Customer ID" is removed from the features vector.

Furthermore, to handle outliers that exist in the dataset, the features are standardized using the Z-score standardization method. This transformation centers the data by subtracting the mean and dividing it by the standard deviation, making it easier to detect and handle outliers based on extreme values relative to the mean and standard deviation.

The data is divided using the holdout method (70/30) for training and testing purposes. The training dataset is split into training and validation datasets using cross-validation techniques with 4 folds. This division helps in model evaluation and hyperparameter tuning. By applying these pre-processing steps, the dataset is prepared for further analysis and modeling, enhancing the effectiveness classification models.

### ii. Hyperparameter optimization

This section focuses on hyperparameter optimization for various classification models like Lasso Logistic Regression, Classification Tree, and Support Vector Machines. Hyperparameters are tuned using a validation dataset, allowing for the selection of optimal parameter values that maximize the model's performance.

For Lasso Logistic Regression, the lambda regularization parameter plays a crucial role. Two common approaches to selecting the optimal lambda are Lambda Minimum Deviance and Lambda 1 SE. Lambda Min Deviance identifies the lambda value that corresponds to the lowest Cross Validated mean squared error. On the other hand, Lambda 1 SE finds a simpler model with an error within one standard error of the best model. The objective is to create a balance between accuracy and simplicity. In this project, both lambda1SE and Lambda Min Deviance suggest a model with 11 features. Although both options are valid, lambda1SE is chosen as the optimal lambda due to its balance between accuracy and model complexity.

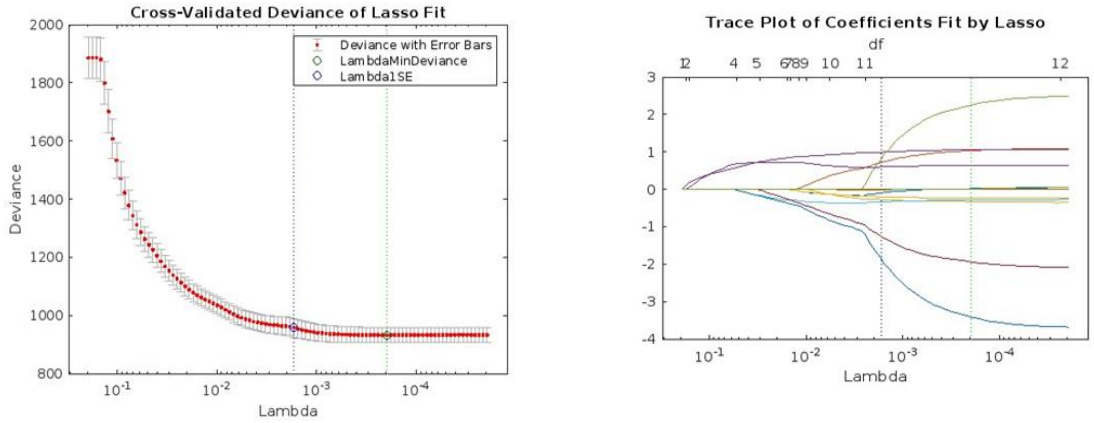


*Figure 6 Cross-Validated Deviance of Lasso Fit and Trace Plot of Coefficients Fit by Lasso*

When dealing with Support Vector Machine (SVM) models, the selection of the appropriate kernel function is essential. The project considers models with various kernel functions, including "Linear", "Polynomial", and “RBF" (Radial Basis Function). The model with the highest cross-validation accuracy is chosen as the optimal SVM model. In this case, the SVM model with the "Polynomial" kernel function demonstrates the highest accuracy among the models.

*Table 3 Selection of Kernel function*

| Kernal Function | Cross Validation Accuracy |
|---|---|
| RBF | 0.9478 |
| Polynomial | 0.9587 |
| Linear | 0.8989 |

For decision trees, determining the minimum number of samples required at a leaf node is crucial. In the project, models with minimum leaf sizes ranging from 10 to 100 were created and evaluated. The model with a minimum leaf size of 10 demonstrates the highest cross-validation accuracy. As a result, the model with a minimum leaf size of 10 is chosen as the optimal decision tree model.

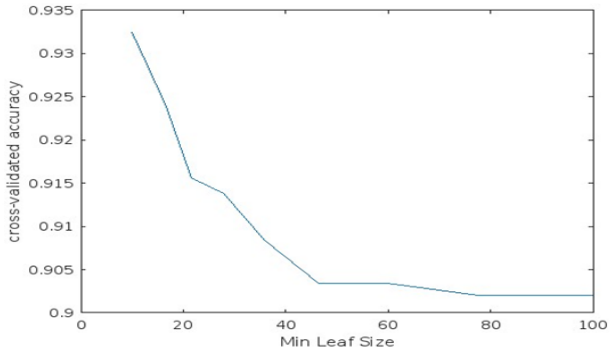


*Figure 7 Decision tree model: minimum leaf size*

By considering cross-validation accuracy and deviance, the project selects the optimal lambda for Lasso Logistic Regression i.e., Lambda1SE, the optimal kernel function for SVM i.e., Polynomial, and the optimal minimum leaf size for the decision tree i.e.,10. These choices help to ensure the model's performance.

### iii. Model Evaluation

This section focuses on evaluating the performance of various classification models using the test dataset, following the optimization of hyperparameters. However, unlike lasso logistic regression, classification tree and SVM model, logistic regression does not require any hyperparameter tuning. Evaluation metrics such as AUC, accuracy, and F1 score are used to compare the models. With this information appropriate choices can be made regarding the most suitable model for the churn dataset.

#### a. AUC and ROC Evaluation

The performance of the classification models can be evaluated using the Area Under the Curve (AUC) metric. The AUC values for Logistic Regression, Lasso Logistic Regression, SVM, and Classification Tree are 0.91, 0.92, 0.96, and 0.95, respectively.

Figure 8 shows that SVM model has the highest AUC value. This suggests that the model has a higher accuracy in predicting churn and non-churn instances. Classification Tree demonstrates a slightly lower AUC value but still performs well in terms of predictive accuracy. The Lasso Logistic Regression model follows Classification tree, indicating reasonable performance in identifying churn instances. Lastly, the Logistic Regression model depicts the lowest AUC among the evaluated models.

These results imply that the SVM and Classification Tree models are more effective in capturing the patterns and underlying relationships within the churn dataset, resulting in better predictive performance. However, other evaluation metrics such as accuracy, precision, recall, and F1 score can help to gain a comprehensive understanding of the model's overall performance and suitability for the churn prediction task.

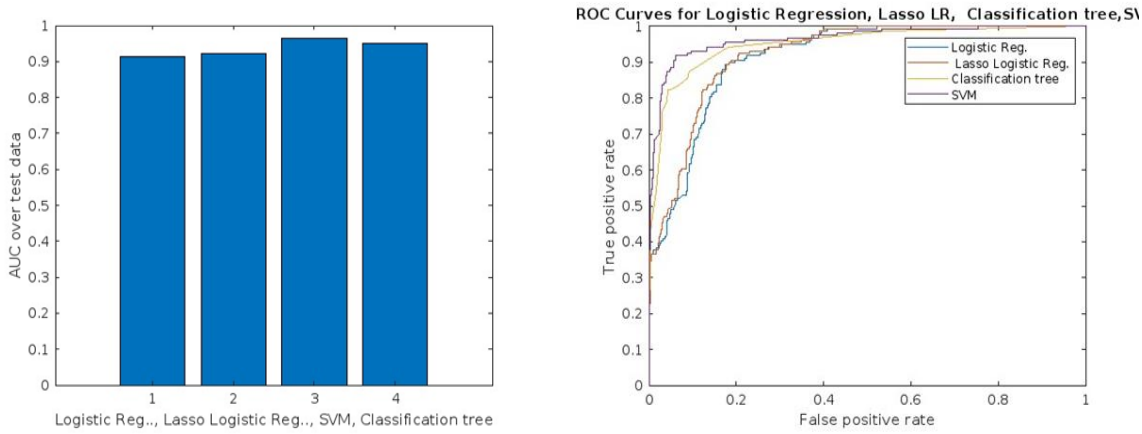


*Figure 8 AUC and ROC for the models*

### b. Evaluation using Confusion Matrix

Table 4 and 5 show that the results obtained from both logistic regression and lasso logistic regression are similar. However, both models had a significant number of false negatives, indicating instances where the models failed to identify actual churn cases correctly resulting in low recall. This suggests that these models may have limitations in accurately predicting churn cases.

The SVM model performed well in terms of accuracy, precision, recall and F1 score. It correctly identified a larger proportion of actual churn cases compared to the other models. However, it also had a higher number of false positive predictions, misclassifying some instances as churn when they were non-churn. This balance between false positives and true positives should be considered based on the specific objectives and requirements of the churn prediction task.

The classification tree model had a balanced performance across all metrics. It correctly classified a good number of non-churn cases and churn cases. However, it also had a significant number of false

positives and false negatives. This suggests that the model may have a moderate level of accuracy in predicting churn cases.

In conclusion, SVM shows the highest overall performance, with a good balance between accuracy, precision, recall, and F1 score. It shows the potential to accurately identify churn cases.

*Table 4 Confusion Matrix*

| Model | True Negative | False Positive | True Positive | False Negative |
|---|---|---|---|---|
| Logistic Regression | 783 | 3 | 58 | 101 |
| Lasso Logistic Regression | 782 | 4 | 58 | 101 |
| SVM | 752 | 34 | 139 | 20 |
| Classification Tree | 753 | 33 | 126 | 33 |

*Table 5 Accuracy and F1 scores*

| Model on test dataset | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Logistic Regression | 0.8899 | 0.9508 | 0.3648 | 0.5273 |
| Lasso Logistic Regression | 0.8889 | 0.9355 | 0.3648 | 0.5249 |
| SVM | 0.9429 | 0.8035 | 0.8742 | 0.8373 |
| Classification Tree | 0.9302 | 0.7925 | 0.7925 | 0.7925 |

## 4. Result Interpretation and discussion

Based on the evaluation metrics including AUC, ROC, accuracy, precision, and recall, the SVM model with the Polynomial kernel function emerged as the best choice as shown in Figure 8, Table 4 and Table 5. This highlights the SVM model's ability to provide a robust and comprehensive prediction of churn instances. The accuracy of the SVM model on test dataset is 94.29% which is close to Jafari-Marandi et al. (2020) who achieved the accuracy of 95% using Artificial Neural Network, KNN and Decision tree. Although this model has high accuracy, this paper is limited to only 3 classification models. Moreover, this paper does not discuss the possibility of feature engineering to improve the model.

Table 2 depicts distinct characteristics between the two clusters. It shows the customers in cluster 1 are more likely to churn. Cluster 1 customers depict lower charge amounts and customer values compared to cluster 2. Furthermore, they tend to lack a contractual traffic plan, indicating a higher likelihood of “pay as you go” customers. Moreover, cluster 1 customers display higher levels of inactivity and less frequent usage of the service for SMS and distinct calls. Although, the customers in cluster 1 experience less call failures and they have more complaints than cluster 2. These findings collectively suggest that cluster 1 represents a segment of customers who are at a higher risk of churn not due to service but their lower engagement and usage patterns.

In conclusion, the model developed in this paper is robust and effective in predicting customer churn. The findings suggest that the company can utilize this model as a valuable tool to identify customers who are likely to churn in the future. The findings from clustering analysis can also be used to verify the predictions from the model. The company can implement effective customer retention plans to target inactive customers with “pay as you go” traffic plan. Furthermore, these strategies may include tailored incentives, improved customer support, enhanced product offerings, or proactive communication to address potential problems and increase customer satisfaction. Therefore, the utilization of this churn prediction model helps the company to take effective measures to retain valuable customers and maintain long-term business growth and success by utilizing the insights gained from the clustering analysis. This would help to develop targeted retention strategies tailored to the specific needs and behaviors of each cluster, ultimately increasing customer satisfaction and reducing churn rates.

## Reference

Jafari-Marandi, R., Denton, J., Idris, A., Smith, B. K., & Keramati, A. (2020). Optimum Profit-Driven

Churn Decision Making: Innovative Artificial Neural Networks in Telecom Industry. Neural Computing and Applications.

THE WORLD BANK (2022). Population, total – Iran, Islamic Rep. Available online: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=IR