Project Report ISyE 6414 A/MSA

Prediction & Analysis of Customer Churn in an Iranian Telecom Company

Group 4

Tyler Burns

Rishidhar Duvvuru

Victor Lai

Rayan Maleck

Spiros Valouxis

Supervised by
Dr. Tokol-Goldsman

Georgia Institute of Technology

December 1, 2023

Contents

1 Introduction

This report tackles the issue of customer churn in a telecom company, exploring ways to predict and address it through regression analysis. Customer churn, or people leaving the service, is a big deal for telecom companies. It's not just about losing money; it also impacts the company's reputation and competitiveness.

Note: The code will be uploaded as a separate file. We will upload the pdf of our R notebook.

1.1 Background

Customer churn is when people drop the telecom service, either by switching to another provider, downgrading their plans, or quitting altogether. High churn rates hurt not only revenue but also the company's image and ability to compete. Traditional methods of retaining customers aren't always cutting it, especially with how fast technology and consumer expectations are changing.

To better understand and tackle customer churn,we're diving into a bunch of data on customer behavior, service use, and demographics. The goal is to find out what factors are driving people to leave, so the company can do something about it.

The report will explain how we did the analysis, what factors we considered, and what we found. The aim is to give the telecom industry practical insights to reduce churn, make customers happier, and stay competitive.

2 Problem Statement

This project tackles the challenge of high customer turnover rates within an Iranian telecom company. The focus is on employing logistic regression techniques to analyze a dataset, with the primary goal of identifying the current causes of customer churn and recommending actionable steps to reduce it. The project also seeks to gain valuable insights into both the business operations and consumer market dynamics.

2.1 Variables

The dataset encompasses various customer-related variables, including profiles, usage patterns, and demographics. Through logistic regression, we aim to discern the key factors influencing churn and pinpoint the most effective variables for prediction.

2.2 Objectives

1. **Churn Pattern Analysis:** Conduct a detailed examination of customer profiles to uncover patterns and trends associated with churn.
2. **Logistic Regression Modeling:** Develop and evaluate logistic regression models to determine the most effective one for predicting churn. This involves identifying the critical variables that significantly contribute to customer turnover.
3. **Insight Generation:** Provide additional insights into both business operations and consumer behavior, facilitating operational optimization and strategic decision-making.
4. **Actionable Recommendations:** Interpret the model results to prescribe concrete, actionable steps for the telecom company to implement. The aim is to empower the company with strategies for reducing churn and enhancing overall customer satisfaction.

By executing these steps, the project seeks to not only address the immediate challenge of customer turnover but also contribute to the telecom company's long-term strategic decision-making process.

3 Data Description and Analysis

The dataset of fourteen columns is composed of 3150 instances randomly collected over a twelve month period. We used the Iranian Telecom Churn dataset from the UCI Machine Learning Repository (source: UCI ML Repository's Iranian Churn Dataset)

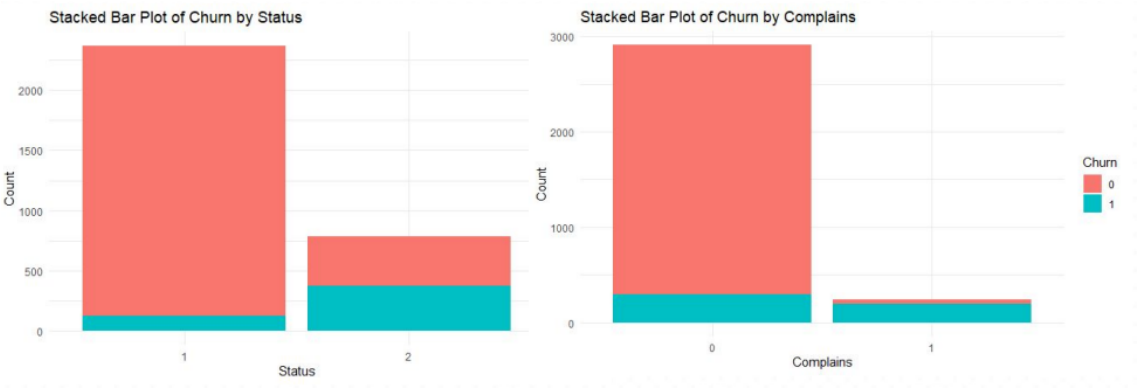

Figure 1: Status - Complaints Barplots

3.1 Dataset Variables

Eight quantitative variables and five qualitative variables will be used to build models to predict the one explanatory variable of Churn. Churn in this project is defined by whether a customer maintains a business relationship with the company or leaves. A customer continuing business is distinguished by a 0, whereas a customer that discontinues business is distinguished with a 1.

In the table below, we use Numerical for Quantitative variables and Categorical for Qualitative variables to make it easier to separate.

Variable Name	**Type**	**Description**
Call Failures	Numerical	Total number of call failures
Subscription Length	Numerical	Total duration of the subscription in months
Distinct Called Numbers	Numerical	Total number of distinct phone calls
Seconds of Use	Numerical	Total duration of all calls in seconds
Frequency of Use	Numerical	Total number of calls
Frequency of SMS	Numerical	Total number of text messages
Customer Value	Numerical	Calculated value of the customer
Age	Numerical	Age of the customer
Complains	Categorical	0 if no complaint submitted, 1 if complaint submitted
Age Group	Categorical	1-5 representing age group from youngest (1) to oldest (5)
Tariff Plan	Categorical	1: pay as you go, 2: contractual
Status	Categorical	1 for active customer, 0 for inactive customer
Charge Amount	Categorical	0-9 representing amount charged from lowest to highest

3.2 Exploratory Analysis

Through bar charts and box plots developed for exploratory analysis, a few conclusions can be made. Inactive customers (status = 2) have an extremely high churn rate compared to active users [Figure:1]. Also, users who submitted complaints had a 50% chance of churning [Figure:1]. Churn appears to decrease when the charge amount increases. Most customers use a pay as you go plan (tariff plan = 1), which is associated with higher levels of churn compared to contractual plans (tariff plan = 2). Customers with the pay as you go plan generally pay significantly less compared to contractual customers. With these discoveries in mind, customer behavior variables are likely to be good predictors of customer churn probability.

The correlation heat map [Figure:2] provides excellent generalized information on the correlation between variables and identifies potential issues with multicollinearity. Usage variables appear to be more correlated with each other, with “frequency of use” and “seconds of use” showing clear signs of multicollinearity. As predicted by the exploratory analysis, customers with complaints and inactive customers are significantly correlated with customer churn. While the matrix heat map is an excellent heuristic, variance inflation factor (VIF) tests are utilized to properly determine multicollinearity (we will explore this isuue further during the model fitting phase).

Note: For more barplots and boxplots regarding our Data Analysis, please visit the Appendix section at the end of the report.

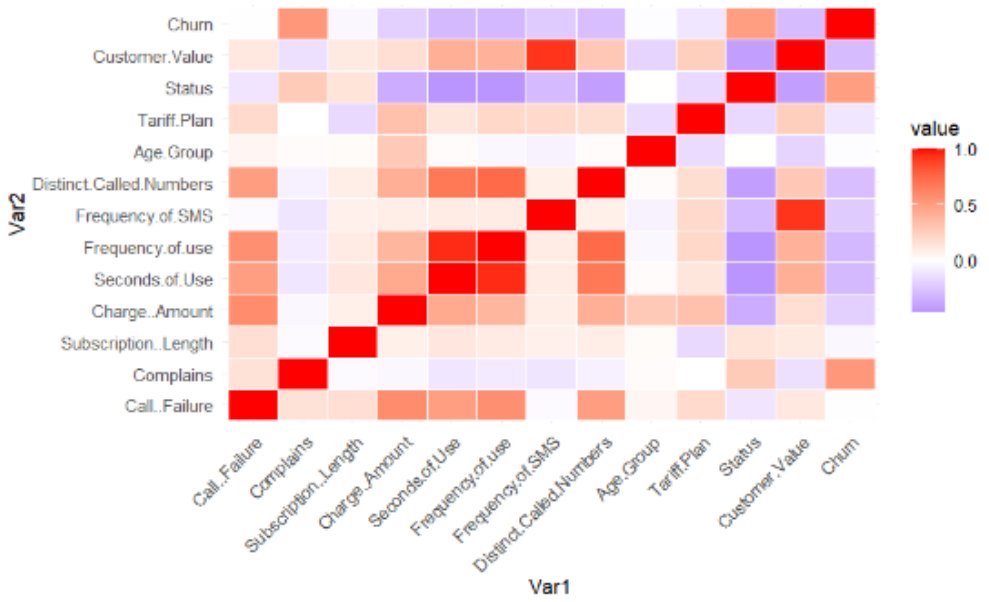

Figure 2: Matrix Correlation Heat Map

4 Model Fitting

For all models, we randomly split the data into train and test data(70/30). We consider a significance level of 0.1 for all models and analyses.

4.1 Full Logistic Regression Model

First, we fit a model to the full dataset with all variables and data points. The model summary [Figure:3] shows that the following variables are significant: Call Failure, Complains, Charge Amount, Frequency of Use, Frequency of SMS, Status, Customer Value. One weird result in this model was that the intercept was not found to be significant.

```
glm(formula = Churn ~ ., family = "binomial", data = train_data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.664e+01  5.014e+02  -0.033  0.97353
Call..Failure            1.653e-01  2.820e-02   5.862 4.58e-09 ***
Complains1               3.887e+00  4.043e-01   9.616  < 2e-16 ***
Subscription..Length    -1.983e-02  1.628e-02  -1.218  0.22314
Charge..Amount          -4.462e-01  1.963e-01  -2.273  0.02300 *
Seconds.of.Use          -7.728e-05  2.598e-04  -0.297  0.76614
Frequency.of.use        -6.425e-02  1.372e-02  -4.683 2.82e-06 ***
Frequency.of.SMS        -7.204e-02  2.201e-02  -3.274  0.00106 **
Distinct.Called.Numbers  2.205e-03  1.387e-02   0.159  0.87366
Age.Group2               1.539e+01  5.014e+02   0.031  0.97552
Age.Group3               1.561e+01  5.014e+02   0.031  0.97517
Age.Group4               1.642e+01  5.014e+02   0.033  0.97387
Age.Group5               1.485e+01  5.014e+02   0.030  0.97637
Tariff.Plan2             7.747e-01  1.001e+00   0.774  0.43897
Status2                  1.275e+00  3.204e-01   3.978 6.94e-05 ***
Customer.Value           1.351e-02  5.011e-03   2.697  0.00700 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1197.4  on 1399  degrees of freedom
Residual deviance:  568.2  on 1384  degrees of freedom
AIC: 600.2
```

Figure 3: Full Model Summary

Next, We knew we needed to test for multicollinearity. We did this by calculating the GVIF of each feature and found that Seconds of Use, Frequency of Use, Frequency of SMS, Customer

Value all had a GVIF greater than 10, indicating some multicollinearity in the model. We knew we needed to remove these in future models.

4.2 Logistic Regression Model - No Multicollinearity

For our next model, we would want to remove the variables causing multicollinearity without losing additional information provided by these variables. So, we created interaction variables like "avg talk time", "Call Failure rate", "Avg calls per number dialed", etc. before removing the correlated variables. After creating the new variables, we checked a new correlation matrix [Figure:4] to see if the updated dataset was less correlated between features. The new matrix shows that the independent variables were less correlated with each other, indicating that multicollinearity is unlikely to be an issue for the new model.

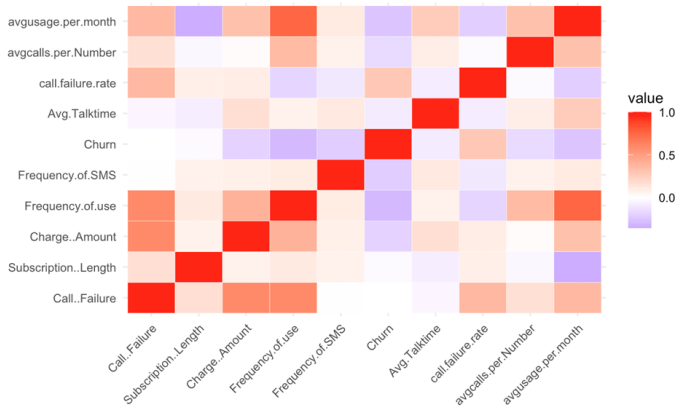

Figure 4: New Variables Correlation Matrix

Next, we had to retrain the model on the updated dataset, and examine the results. We found that Call Failure, Complains, Charge Amount, Frequency of Use, Frequency of SMS, Status, Customer Value, Avg calls per number, Avg usage per month were all significant variables in the model. We still had the statistically insignificant intercept in this model. We still needed to check this model for multicollinearity so we can test whether or not our method to remove it worked. As shown in [Figure:5], each of the features in this model had a GVIF of less than 10, indicating that we successfully removed multicollinearity from our model.

	GVIF	Df	GVIF^(1/(2*Df))
Call..Failure	3.353955	1	1.831381
Complains	1.171691	1	1.082447
Subscription..Length	1.941790	1	1.393481
Charge..Amount	2.428447	1	1.558347
Frequency.of.use	3.454169	1	1.858539
Frequency.of.SMS	2.049576	1	1.431634
Age.Group	1.593310	4	1.059955
Tariff.Plan	1.859141	1	1.363503
Status	1.959336	1	1.399763
Avg.Talktime	1.258479	1	1.121819
call.failure.rate	1.848863	1	1.359729
avgcalls.per.Number	1.281503	1	1.132035
avgusage.per.month	3.113549	1	1.764525

Figure 5: Second Model Multicollinearity Test

4.3 Logistic Regression Model - No Outliers

Next, we wanted to do some outlier analysis to see if any outliers were hurting our model's performance. By calculating the cook's distance of each point and plotting them [Figure:6], we found 4 outliers. One thing to note is that with so many data points in our dataset, the cook's distance values are deflated. Then, we retrained the model on our data without the outlier points we found.

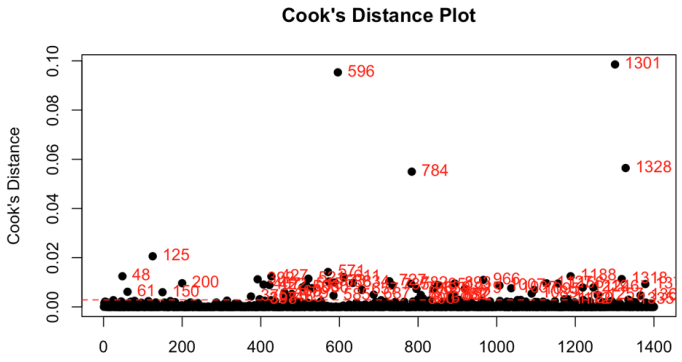

Figure 6: Cook's Test

Examining the results [Figure:7] shows that removing the outliers has slightly improved the significance of some of the variables and reduced the AIC value. However, the intercept is still not statistically significant. Unfortunately, this model produced an error. The warning "glm.fit: fitted

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = Churn ~ ., family = "binomial", data = train_no_outliers)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.807e+01  1.311e+03  -0.014 0.989004
Call..Failure         1.516e-01  3.288e-02   4.611 4.02e-06 ***
Complains1            3.648e+00  4.074e-01   8.953  < 2e-16 ***
Subscription..Length  2.403e-03  1.861e-02   0.129 0.897252
Charge..Amount       -4.485e-01  1.622e-01  -2.766 0.005679 **
Frequency.of.use     -3.728e-02  6.785e-03  -5.495 3.90e-08 ***
Frequency.of.SMS     -4.074e-02  8.716e-03  -4.674 2.95e-06 ***
Age.Group2            1.652e+01  1.311e+03   0.013 0.989949
Age.Group3            1.674e+01  1.311e+03   0.013 0.989814
Age.Group4            1.717e+01  1.311e+03   0.013 0.989551
Age.Group5            1.438e+01  1.311e+03   0.011 0.991250
Tariff.Plan2         -1.393e+01  7.172e+02  -0.019 0.984498
Status2               1.195e+00  3.239e-01   3.690 0.000224 ***
Avg.Talktime         -1.983e-04  2.634e-03  -0.075 0.939991
call.failure.rate     7.754e-01  6.028e-01   1.286 0.198299
avgcalls.per.Number  -2.328e-01  7.891e-02  -2.950 0.003179 **
avgusage.per.month    7.890e-03  1.798e-03   4.387 1.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.28  on 1395  degrees of freedom
Residual deviance:  539.67  on 1379  degrees of freedom
AIC: 573.67
```

Figure 7: No Outliers Model Summary

probabilities numerically 0 or 1 occurred" in logistic regression often indicates that the model is having difficulty estimating probabilities for extreme values of the predictors. Therefore, removing influential points doesn't seem to be the right approach.

4.4 Logistic Regression Model - Stepwise

Next, to decrease model complexity, we applied stepwise regression to eliminate potential insignificant variables and overfitting from the model. By examining our results [Figure:8], we found that the intercept is now statistically significant. Predictors like Status 2 (inactive), Complains 1(Yes), and Call Failures increase the odds of customer churn per unit increase in the respective predictors while others are kept constant. Predictors like Frequency of use, SMS, Charge Amount, and Avg calls per number decreases the odds of customer churn per unit increase in the respective predictors while others are constant. Because the results of our stepwise model seemed promising, we wanted to test for overall significance and goodness of fit. For significance, we performed the chi-squared test, and we calculated a p-value of 0. So, we reject the null hypothesis and conclude that the model is statistically significant. For our goodness of fit test, we tested on the deviance and Pearson residuals. For each, we want a high p-value, and we found a p-value of 1.0. Thus, we conclude that the model fits the data well. Through our early analysis of our stepwise model, it appears to be a potentially excellent choice for our problem.

```
Call:
glm(formula = Churn ~ Status + Complains + Frequency.of.use +
    Call..Failure + Frequency.of.SMS + Charge..Amount + avgusage.per.month +
    avgcalls.per.Number + call.failure.rate, family = "binomial",
    data = train_data2)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.580629   0.295678  -5.346  9.0e-08 ***
Status2             1.149371   0.278587   4.126  3.7e-05 ***
Complains1          3.914692   0.412679   9.486  < 2e-16 ***
Frequency.of.use   -0.035566   0.006269  -5.674  1.4e-08 ***
Call..Failure       0.119163   0.029891   3.987  6.7e-05 ***
Frequency.of.SMS   -0.017636   0.005367  -3.286 0.001017 **
Charge..Amount     -0.503015   0.139732  -3.600 0.000318 ***
avgusage.per.month  0.004351   0.001210   3.594 0.000325 ***
avgcalls.per.Number -0.104702   0.060499  -1.731 0.083515 .
call.failure.rate   0.828733   0.579616   1.430 0.152775
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1197.37  on 1399  degrees of freedom
Residual deviance:  580.44  on 1390  degrees of freedom
AIC: 600.44
```

Figure 8: Stepwise Model Summary

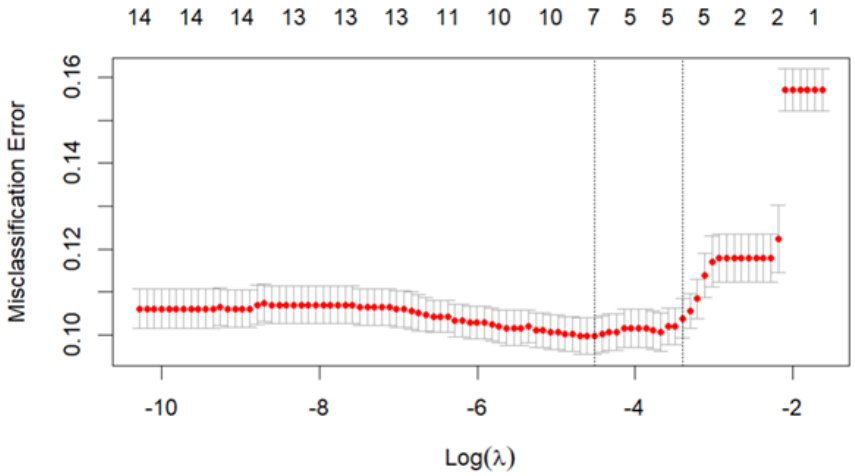

Figure 9: LASSO Regularization Plot

4.5 Logistic Regression Model - LASSO

We wanted to continue trying some more models to ensure we made the best model choice. The next one we tried was a logistic regression using lasso regularization. We decided to try this because of the multicollinearity we detected in earlier models. Lasso models have a penalization term that when being optimized reduces some coefficients to 0. This mitigates any multicollinearity concerns from our model.

We used the glmnet function in R to train a lasso model and printed [Figure:9]. The regularization increases the further right on the x-axis the data point is. The first vertical line is the lambda that produces the lowest error. We selected this lambda and viewed the coefficients [Figure:10]. The selected coefficients were complains, subscription length, frequency of use, distinct called numbers, age group 5, status, and customer value.

Next, we calculated the optimal threshold for this model, which was calculated to be 0.27, using the same technique as previous models.

Finally, we needed to run a goodness of fit test to ensure that the model fit the data well. Because the glmnet function does not have a built-in deviance, we decided to use the Hosmer-Lemeshow goodness of fit test. We found a p-value of 0.0003, which indicates that the model does not fit the data well. Thus, we decided not to explore our lasso model further in our validation step.

```
16 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)             -1.669576450
(Intercept)              .
Complains1               3.433064682
Subscription..Length    -0.005773390
Charge..Amount           .
Seconds.of.Use           .
Frequency.of.use        -0.005617937
Frequency.of.SMS         .
Distinct.Called.Numbers -0.013847857
Age.Group2               .
Age.Group3               .
Age.Group4               .
Age.Group5              -0.452452147
Tariff.Plan2             .
Status2                  1.476588345
Customer.Value          -0.001104077
```

Figure 10: LASSO Coefficients

4.6 Random Forest

We decided it might be interesting to have a more complex model to compare our results with, so we chose to create a random forest model. We used 1000 trees and a node size of 5 for our model. Our results can be seen in Figure:11.

```
                         Sample size: 2204
                    Number of trees: 1000
          Forest terminal node size: 5
      Average no. of terminal nodes: 76.737
No. of variables tried at each split: 5
             Total no. of variables: 15
      Resampling used to grow trees: swor
   Resample size used to grow trees: 1393
                           Analysis: RF-R
                             Family: regr
                      Splitting rule: mse *random*
      Number of random split points: 10
                    (OOB) R squared: 0.72008087
  (OOB) Requested performance error: 0.03706195
```

Figure 11: Random Forest Model Results

Our model had an R2 of 0.72, which indicates the model does decently well at explaining variability in our y. It had an error rate of 3.7%, which is excellent. However, our goal for this project is not just to get the most accurate model. We also want to be able to see which variables are the most important. To achieve this, we produced a table of the variable importance using the VIMP function in R [Figure:12]. The 4 most important variables are complains, status, subscription length, and seconds of use.

Variable Importance

	x
X.Intercept.	0.0000000
Complains1	0.3639560
Subscription..Length	0.0800021
Charge..Amount	0.0096145
Seconds.of.Use	0.0688263
Frequency.of.use	0.0546812
Frequency.of.SMS	0.0196713
Distinct.Called.Numbers	0.0460106
Age.Group2	0.0217788
Age.Group3	0.0078184
Age.Group4	0.0179016
Age.Group5	0.0033748
Tariff.Plan2	0.0017470
Status2	0.1745406
Customer.Value	0.0251976

Figure 12: Random Forest Variable Importance

5 Validation

In the course of developing and fine-tuning our logistic regression models for customer churn prediction in the telecom industry, the determination of an optimal threshold is a crucial step in balancing the trade-off between different evaluation metrics. In particular, we have prioritized two key metrics: recall and accuracy.

5.1 Choosing the Threshold

In consideration of the business context, where the cost of acquiring a new customer is significantly higher than retaining an existing one, we emphasize the importance of minimizing false negatives (churn cases incorrectly classified as non-churn). This aligns with our goal to proactively identify and retain customers who may be at risk of churning. After extensive experimentation and validation, we have identified the optimal threshold for our stepwise model as 0.18 [Figure:13]. At this threshold, we strike a balance that maximizes recall (true positive rate) while maintaining a satisfactory level of accuracy.

It's important to acknowledge that the chosen threshold is context-specific and hinges on the cost implications of customer acquisition versus retention. Our emphasis on recall is a strategic decision to minimize the false negatives, ensuring that we identify as many potential churn cases as possible. If the cost of acquiring customers were not as substantial, we recognize that an alternative threshold of 0.37 could be considered. At this threshold, the model may exhibit a higher precision (lower false positive rate) at the expense of some recall. This would be a more suitable choice in scenarios where the cost of false positives is more manageable.

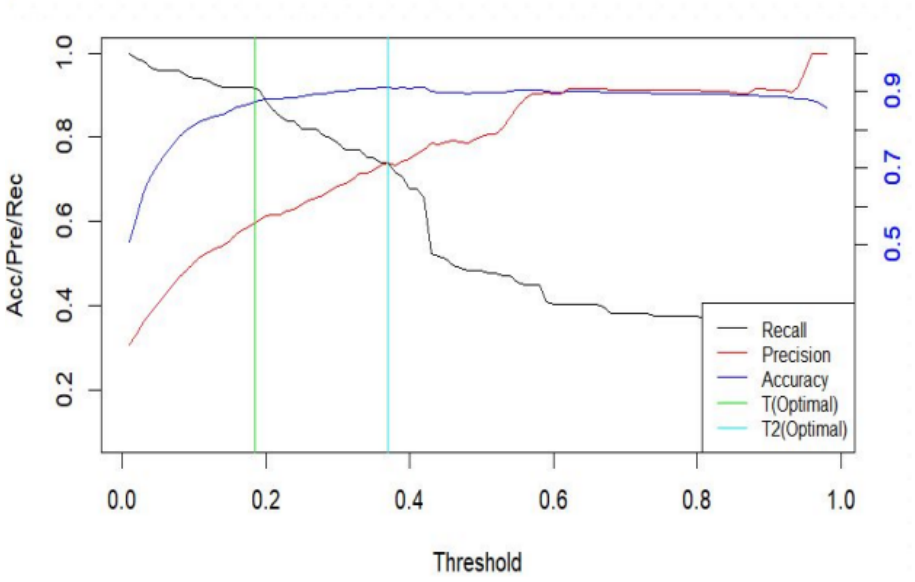

Figure 13: Stepwise Model Threshold Selection

In conclusion, the threshold selection process is a critical aspect of optimizing our logistic regression models for customer churn prediction. Our approach reflects a careful consideration of the business landscape, where the balance between acquiring and retaining customers plays a pivotal role.

Moving forward, continuous monitoring and reevaluation of model performance, alongside potential adjustments to the threshold based on evolving business dynamics, will be integral to maintaining the efficacy of our customer churn prediction system.

5.2 ROC Curves

ROC (Receiver Operating Characteristic) curves are graphical representations that illustrate the performance of a classification model across different thresholds. They depict the trade-off between Sensitivity (True Positive Rate) and Specificity (True Negative Rate). The diagonal line in the ROC space represents the performance of a random classifier, while curves above the diagonal indicate better-than-random performance [Figure:14].

In our analysis of ROC curves for each model, we observe the intricate balance between Sensitivity and Specificity. Minor differences exist in the performance of the models, indicating comparable predictive capabilities. The area under the ROC curve (AUC-ROC) provides a summary measure of overall model performance. We note that all models exhibit AUC values indicative of reasonable discriminative ability.

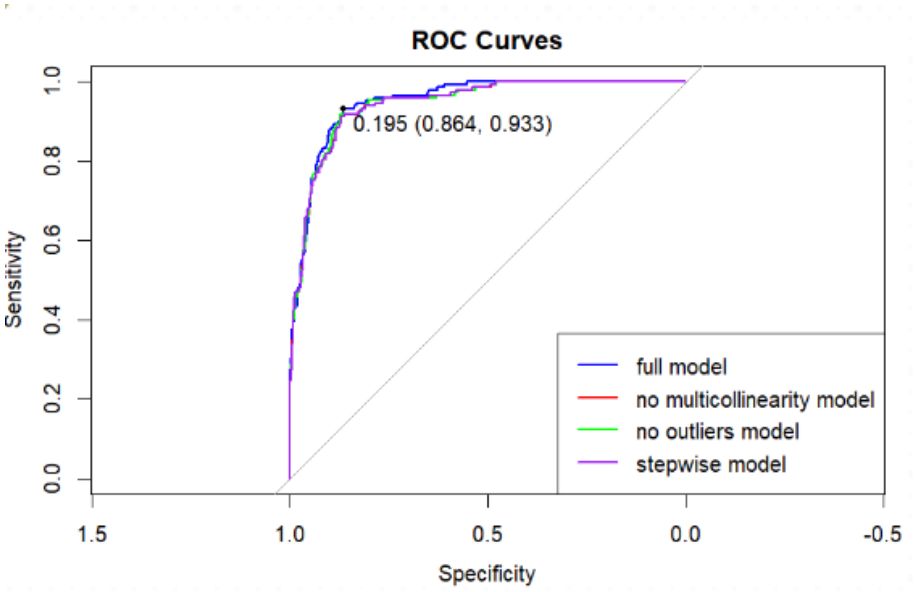

Figure 14: ROC Curves of the 4 models

5.3 Model Comparison

Now, let's delve into the detailed results of our model comparison based on the test data. The table below summarizes the key performance metrics for each model, showcasing their validation thresholds, accuracy, precision, recall, F1-score, and additional indicators of importance and goodness of fit.

Model	**Threshold**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Full Model	0.20	0.878	0.571	0.919	0.704
No Collinearity Model	0.19	0.879	0.573	0.919	0.706
No Outliers Model	0.19	0.879	0.573	0.919	0.706
Stepwise Model	0.18	0.876	0.567	0.913	0.70
RF Model (Extra)	0.32	0.959	0.831	0.926	0.88

In light of the minor performance differences and the need for a model that balances interpretability and generalization, we opt for simplicity and robustness. The stepwise model emerges as the preferred choice due to its interpretability, likelihood of generalizing well to new data, reduced risk of overfitting, and computational efficiency.

Notably, the Random Forest model demonstrates a departure in performance from the other models. We built this model as something extra (we did not talk about it during class) and we choose to select a more interpretable model as our final choice.

The decision to choose the **stepwise model as our final model** is grounded in a strategic trade-off between performance and simplicity. A simpler and more interpretable model not only aligns with our business goals but also enhances the model's utility for stakeholders who may not be well-versed in complex model architectures. The stepwise model, with its emphasis on key features and streamlined complexity, is poised to provide actionable insights and facilitate informed decision-making in the context of customer churn prediction.

6 Conclusion

6.1 Summary of findings

Our project addressed the customer churn of an Iranian telecom company by applying various logistic regression techniques to unravel the pivotal factors influencing churn. In order to make our final analysis, we chose the Stepwise Logistic Regression model for our final analysis due to its simplicity, precision, and interpretability. Additionally, the stepwise logistic regression model is most likely to generalize well with new data, thereby reducing the risk of overfitting and maintaining computational efficiency.
This model identified complaints, customer status, charge amount, and call failures as significant predictors of churn, as outlined in the subsequent table depicting the evolving odds linked to each coefficient. Notably, complains emerged as the most influential factor, increasing the odds of churn by 50 times. Additionally, Inactive customer are 3.15 times more likely to leave, while an increase in charge amount yielded a 40% reduction in the likelihood of churn. Each call failure, on the other hand, increased the odds of churn by 1.13 times.

Feature	Odds Ratio
Status	0.206
Complaint	3.156
Frequency of use	0.965
Call Failure	1.127
Frequency of SMS	0.983
Charge Amount	0.605
avgusage per month	1.004
avgcalls per number	0.901
Call failure rate	2.290

6.2 Implications for the company

These findings offer strategic insight into curbing customer churn. Primarily, the exponential increase in churn odds associated with complaints demands immediate action. Swift responses, potentially coupled with commercial gestures or incentives, are imperative to retain customers facing grievances, considering their heightened likelihood of departure.
The relationship observed between higher-priced plans and reduced churn rates indicates that customers perceive enhanced value in these premium plans. Aligning the quality and attractiveness of services with distinct pricing tiers could significantly bolster customer retention. Hence, assisting customers in selecting plans suited to their usage patterns is crucial, as some may initially underestimate their needs, leading to inappropriate plan selection. Initiatives encouraging greater service utilization among customers could also contribute to improved retention rates as inactive users are much more likely to leave.
Additionally, investigating repeated call failures is critical; while a single failure might seem insignificant, their cumulative impact could prove detrimental. Addressing these issues promptly, perhaps through commercial gestures, can effectively mitigate heightened churn probabilities caused by a high number of call failures.

6.3 Recommendations for future endeavors

To bolster the precision and depth of our study, avenues for expansion are abundant. Incorporating additional predictors not present in the current dataset, such as regional data, competitor information, detailed plan specifications, device insights, and usage patterns, could yield a more comprehensive understanding. Exploring alternative models like Decision Trees, Support Vector Machines, K-Nearest-Neighbors, and deep learning methods could augment predictive capabilities. Furthermore, temporal insights through Times Series Analysis would unveil trends, seasonal patterns, and other time-based churn influencers. Lastly, expanding the dataset with more data points could empower the model to discern intricate patterns and fortify predictive accuracy, especially if the current dataset's limitations constrain variable consideration.

7 Appendix

In this section, we present additional observations and insights from the exploratory data analysis (EDA) that complement the main findings discussed in the report.

Charges Barplots [Figure 15]: This barplot visualizes the distribution of charges among customers. It provides insights into the range and distribution of charges, helping to identify patterns in customer spending.

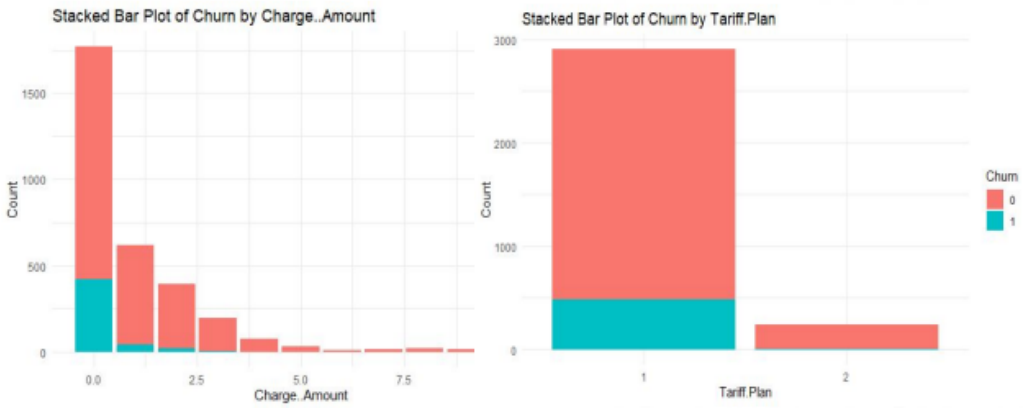

Figure 15: Charges Barplots

Charges Boxplot [Figure 16]: The boxplot offers a detailed view of the distribution of charges by plan, highlighting key statistics such as median, quartiles, and potential outliers. It aids in understanding the variability in customer plans.

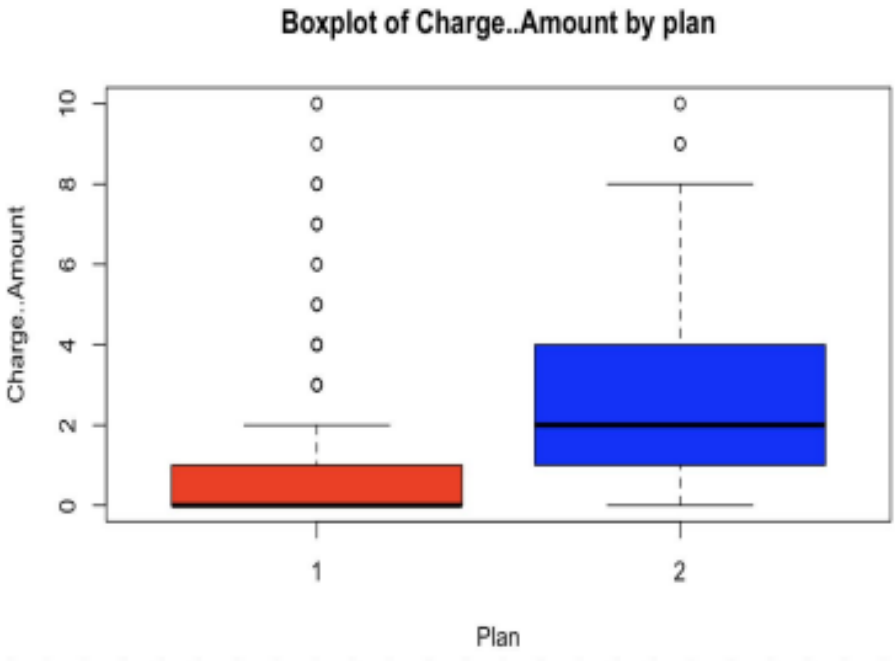

Figure 16: Charges Boxplot

Status and Complaints Barplots [Figure 17]: These barplots depict the relationship between customer status (active or inactive) and the submission of complaints. It helps in assessing whether there is a correlation between customer activity, complaints, and churn.

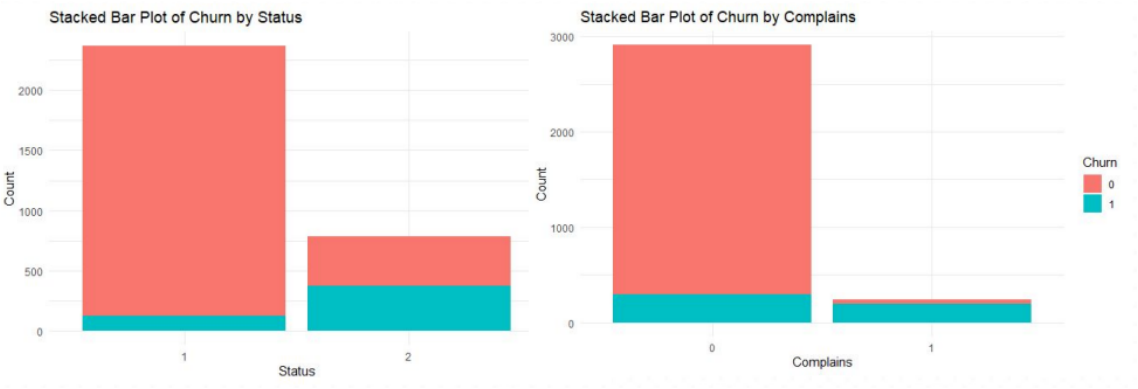

Figure 17: Status and Complaints Barplots

Usage Boxplots [Figure 18]: The boxplots for usage variables, such as call failures, subscription length, and others, provide a visual summary of their distribution. This aids in identifying potential patterns or outliers in customer usage behavior.

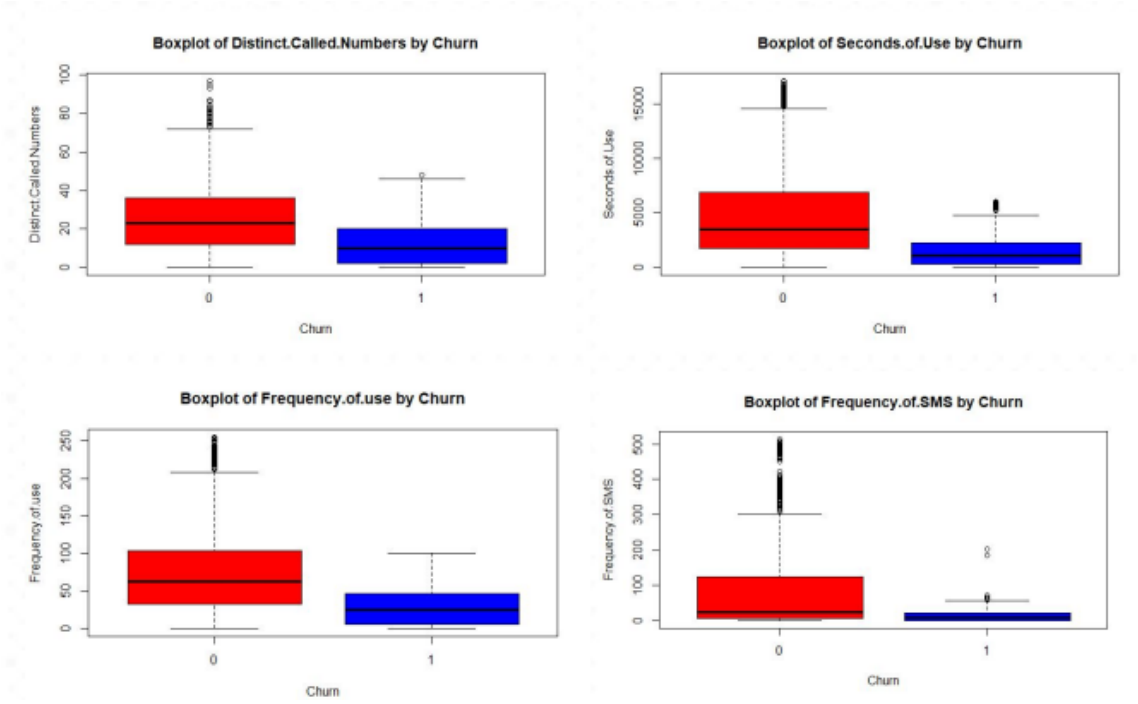

Figure 18: Usage Boxplots

Call Failure & Subscription Length Boxplots [Figure 19]: These boxplots illustrate the connection between call failures, subscription length and churn.

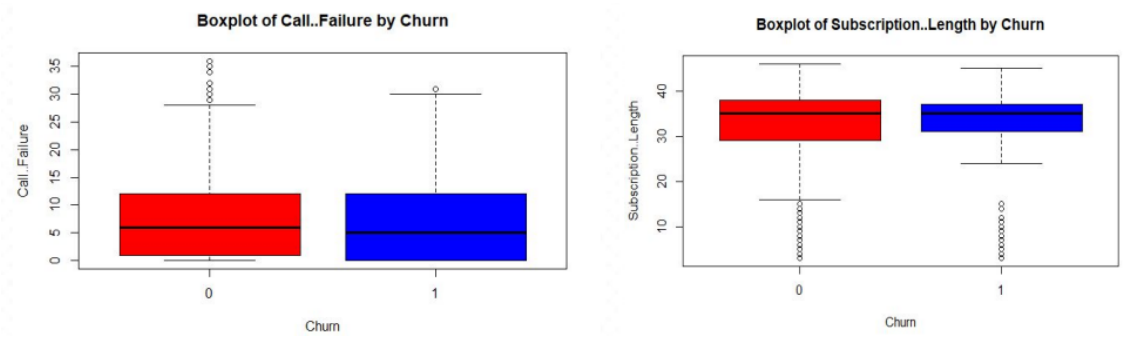

Figure 19: Call Failure & Subscription Length Barplots

General VIF [Figure 20]: Here, we can see some multicollinearity issues we faced before transforming some of the variables.

	GVIF	Df	GVIF^(1/(2*Df))
Call..Failure	2.587487	1	1.608567
Complains	1.155179	1	1.074793
Subscription..Length	1.317625	1	1.147879
Charge..Amount	2.481765	1	1.575362
Seconds.of.Use	30.197970	1	5.495268
Frequency.of.use	16.672830	1	4.083238
Frequency.of.SMS	30.605615	1	5.532234
Distinct.Called.Numbers	2.614427	1	1.616919
Age.Group	2.235154	4	1.105767
Tariff.Plan	1.394949	1	1.181080
Status	1.957999	1	1.399285
Customer.Value	63.126309	1	7.945207

Figure 20: General VIF

Note: The code will be uploaded as a separate file. We will upload the pdf of our R notebook.