UCI Machine Learning Repository - Iranian Churn Final Report

IDS 702 Group 6

Gunel Aghakishiyeva, Barbara Flores, Gavin Li, Tianji Rao

Table of contents

1 Abstract

The purpose of this research was to understand customer behavior in an Iranian telecommunications company. Specifically, two key questions were addressed: 1) Which variables impact customer value? and 2) ? We utilized a dataset encompassing 3,150 customers and 14 variables. The first question was addressed using Linear Regression for Customer Value, and the second with Logistic Regression for Churn prediction, given the continuous and binary nature of the respective variables. Variable selection was done a priori, based on gathered industry knowledge.
Addressing the first research question regarding variables impacting customer value, our analysis unveiled several key factors: Subscription Length, Complaints about Call Failures, Charge Amount, Frequency of SMS, and the choice of a Contractual Plan. Notably, the interaction between a High Charge Amount and a Contractual Plan exhibited a considerable effect on customer value. By employing Linear Regression, these findings provide a nuanced understanding of the drivers of customer satisfaction and loyalty within the telecommunications industry.
Turning to the second research question, focused on the effective prediction of customer churn, Logistic Regression played a fundamental role. The model identified that customers expressing dissatisfaction through complaints had a significantly higher probability of cancellation, with each additional call failure increasing the likelihood of cancellation. Introducing the concept of "User Stickiness," the study highlighted that longer subscription periods correlated with a lower probability of cancellation. The prediction model demonstrated its effectiveness in anticipating customer cancellations, providing practical insights for industry decision-makers.

2 Introduction

The objective of this study is to address two pivotal questions that hold significant implications for the telecommunications industry:

Question	Outcome	Type
Which variables have impact on customer value?	**Customer Value**	**Continuous**
How can we forecast customer churn effectively?	**Churn**	**Binary**

In our study, we analyze the Iranian Churn Dataset from the UCI Machine Learning Repository (UCI, 2020), comprising **3,150 rows** and **14 variables** over 12 months from an Iranian telecom company, focuses on customer attributes, to understand customer value and predict churn in the telecommunications industry. Churn, a term referring to customers discontinuing their service, is a vital metric impacting company revenues and market stability. Customer value, a term represents the total revenue a company has gained from an individual customer over a certain period. This research is crucial for telecom companies to enhance customer satisfaction, reduce churn, and stay competitive. By identifying the key variables affecting customer value and churn, we provide actionable insights for strategic decision-making. This understanding not only aids in retaining valuable customers but also in improving overall customer experience, service quality, and marketing effectiveness, crucial for navigating the competitive telecom landscape.

3 Methods

3.1 Data Processing

Regarding the original database, no missing values were found, so there was no need for handling them. But, some categorical variables that originally had numerous categories were consolidated into simpler categories. This process was carried out with the aim of addressing the complexity inherent in a large number of levels, thus enabling a "Reduction of complexity" in the model. The resulting simplification enhances the interpretation of results, making the analysis more accessible. The following transformations were applied to the dataset:

1. The variables Complains, Age Group, Tariff Plan, Status, and Churn, despite being numeric, were converted into categorical variables and transformed into factor type due to their nature.

2. Due to the variable Charge Amount having 11 levels, a binary variable High Charge was created, indicating whether Charge Amount is considered high or not. This was done to use the variable to generate an interaction term later with Tariff Plan. Otherwise, we would have more than 20 categories when analyzing the interaction.
3. Similar to the previous case, the variable 'Call Failure Indicator was created, which is a binary representation of the numerical variable Call Failure, indicating whether there were call failures or not. This was done to calculate the interaction term between whether a customer had Call Failure or not and whether they had Complains.
4. The interaction variable Tariff Plan High Charge Interaction was constructed to explore the relationship between the Tariff Plan and the binary indicator High Charge. The interaction term considers the combined influence of the tariff plan and the high charge condition on the Customer Value.
5. Similar to the previous case, an interaction variable, Call Failure Complains Interaction, was constructed using the variables Call Failure Indicator and Complains. This interaction variable captures the combined impact of call failures and customer complaints on the likelihood of Churn.

3.2 *Which variables have impact on customer value?*

3.2.1 Exploratory Data Analysis

After cleaning the database, an exploratory data analysis was conducted. Initially, a histogram was created to understand the distribution of the customer value variable. Subsequently, this variable was plotted against certain factors that, according to our industry knowledge, should influence this dependent variable. The objective was to examine whether there was any discernible pattern or relationship between these independent variables and customer value.

3.2.2 Model

We employed a linear regression model to examine our dataset, We ensured robust analysis by identifying andevaluating influential points through Cook's Distance. Additionally, we're safeguarding the model's integrity byrigorously checking for multicollinearity, utilizing both the correlation matrix and the Variance Inflation Factor (VIF) for each predictor to prevent any distortion in our finding.

3.2.3 Model Assessment

To validate the assumptions critical to our linear regression model, we will employ scatterplots and residual plots to confirm **linearity**, scrutinize the study's design to ensure the **independence of errors**, utilize Quantile-Quantile (Q-Q) plots to examine the **normality of errors**, and analyze residual plots for patterns that might indicate violations of the **equal variance of errors assumption**. These checks are vital to the credibility and accuracy of our model's predictions.
In the assessment of our linear regression model, a key metric of focus will be the *adjusted* R^2 value. In the case of our research question and to understand which variables and to what extent they truly impact customer value, we focused on analyzing the coefficients of the variables and their significance.

3.3 *How can we forecast customer churn effectively?*

3.3.1 Exploratory Data Analysis

Similar to the previous research question, after cleaning the data, we proceeded to conduct an exploratory analysis to understand how the churn variable behaves, its distribution within our dataset, and then proceeded to plot the variable against some potentially predictive variables using boxplots. This approach allowed us to visually assess any discernible patterns or trends in the relationship between the churn variable and the selected predictor variables, providing valuable insights into potential factors influencing customer churn within our study.

3.3.2 Model

In our study, we use a logistic regression model to analyze the Churn outcome in the dataset. This model is ideal for our binary response variable, Churn, as it effectively predicts the probability of customer churn based on various predictors, offering critical insights into customer retention factors.

3.3.3 Model Assessment

Influential points

To handle potential outliers, Cook's distance was utilized, and values that deviated significantly from the overall trend of the model were removed. The removal of these outliers was done with the aim of enhancing the robustness and validity of the regression model.

Multicollinearity

To maintain the validity of our logistic regression model, we will examine multicollinearity. This will involve analyzing a correlation matrix and computing the Variance Inflation Factor (VIF) for each predictor to detect and mitigate any interdependencies between variables.
We also find that frenquecy and seconds of use are highly correlated, which is consistent with the first research result. So we decide to drop seconds of use from our model to mitigate the multicollinearity.

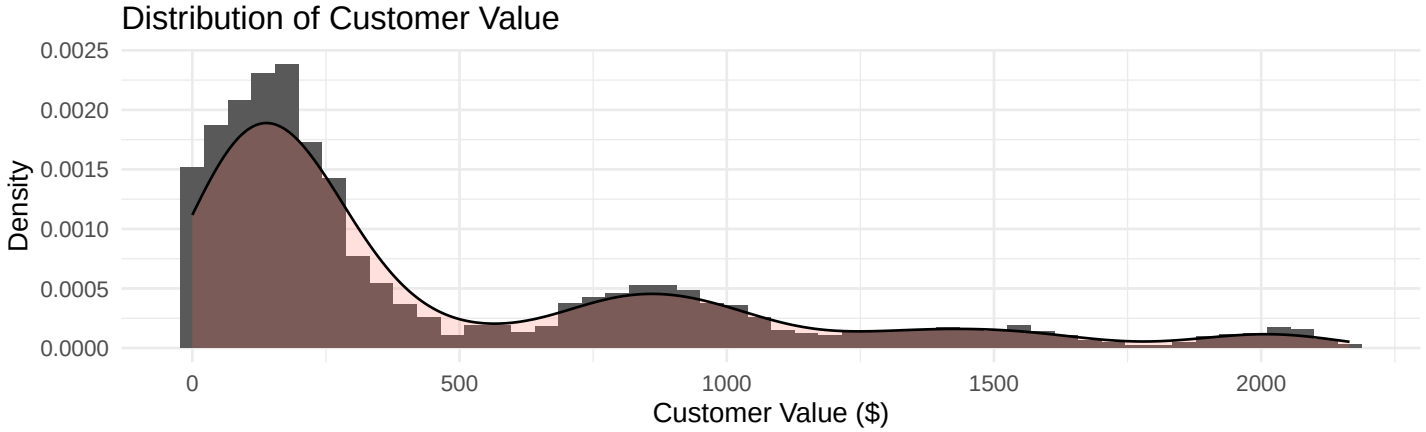
Variable Selection and Cross-validation

A further variable selection is also conducted aims to boosting the predictive power. With the consideration of out-of-sample predictive ability, we can run K-fold Cross Validation on this regression. In this project, we want to test if the traiff plan can help with forecasting, so we import a 5-fold Cross Validation. Here we mainly focus on accuracy and kappa as criterins.
In evaluating our logistic regression model, we will focus on the accuracy, kappa, and Area Under the Receiver Operating Characteristic (ROC) Curve, commonly known as the AUC-ROC.
Accuracy can give us a straightforward understanding of our predictive model performance, while kappa is more useful when dealing with imbalanced datasets like our customer churn dataset. The AUC-ROC measures the model's ability to discriminate between the two classes (churn and no churn) and is more informative than mere accuracy in imbalanced datasets. A higher AUC-ROC value indicates better model performance.

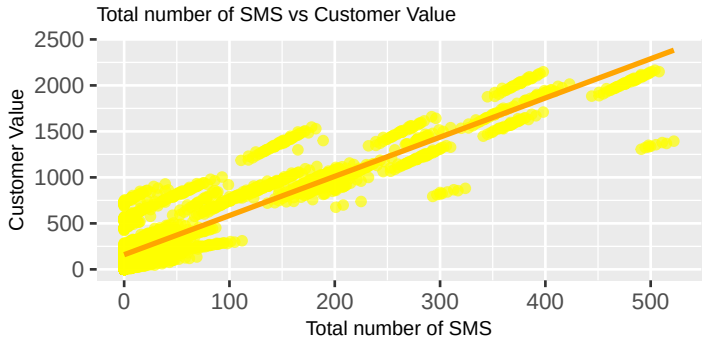
4 Results

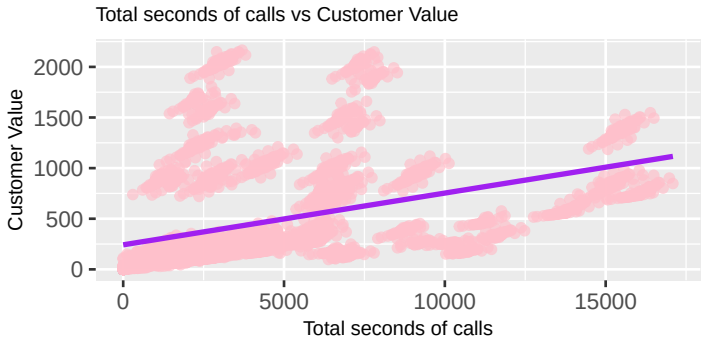
4.1 *Which variables have impact on customer value?*

4.1.1 Exploratory Data Analysis



In the histogram above, we can observe the distribution of the variable customer value within the customer universe.
In the next graphs, we can see how the variable customer value is directly related to the variables Total Number of SMS and Total seconds of call. This signals to us that we could consider these variables in the customer value prediction model.





4.1.2 Model

We first developed a baseline model, incorporating predictors: Call Failure, Charge Amount, Seconds of Use, Frequency of Use, Frequency of SMS, Tariff Plan, and an interaction term between Tariff Plan and Charge Amount.

Residual plots

After reviewing the residual plot, it has become evident that the residuals do not display constant variance across the range of fitted values (Please refer to Appendix 7.2). Such heteroscedasticity suggests that our model could be improved by applying a **log transformation**.

Influential points & Multicollinearity

We evaluated potential influential points using Cook's Distance. The results were clear: no significant influential points were detected, thus ensuring the robustness of our findings. Moving forward, after examination of the correlation matrix and Variance Inflation Factor (VIF) scores in our base linear regression model, we identified a collinearity issue between Seconds of Use and Frequency of Use. To enhance our model's accuracy and interpretability, we decided to exclude Frequency of Use. This action effectively reduces redundancy in our model without significantly compromising its integrity or the validity of our conclusions.
Subsequently, we developed an updated linear regression model, aptly named 'linear2'. In this iteration, we implemented a log transformation of the response variable to correct for heteroscedasticity. To tackle the issue of multicollinearity, we removed 'Frequency of Use' from our predictors. Additionally, we streamlined 'Charge Amount' from a complex ten-category variable to a more manageable two-category one. These refinements serve to simplify the model, enhancing its focus and interpretability, while still maintaining its robust explanatory power.

4.1.3 Model Assessment

Assumption checks

Our assessment of the linear regression model's assumptions has yielded mostly positive results. We've established linearity through and residual plots, confirmed the independence of errors from the study's design, and ascertained equal variance of errors with no significant patterns detected in the residual plots. However, it appears that the assumption of the normality of errors warrants closer attention, as the Quantile-Quantile (Q-Q) plots suggest a slight deviation from the expected distribution. This finding will be given due consideration to ensure the overall validity of our model's conclusions.

Model Summary

Below is a summary table displaying the coefficients, standard error, and p-value for our updated model.

Coefficient	Estimate	Std. Error	Pr(>|t|)
	2.495e+00	2.931e-01	< 2e-16 ***
Subscription Length	-1.327e-02	8.601e-03	0.12286
High Charge Amount	8.957e-01	1.636e-01	4.69e-08 ***
Seconds of Use	3.162e-04	1.858e-05	< 2e-16 ***
Frequency of SMS	1.086e-02	6.583e-04	< 2e-16 ***
Contractual Plan	1.928e+00	6.159e-01	0.00176 **
High Charge Amount : Contractual Plan	-2.108e+00	6.800e-01	0.00195 **

Model Interpretation

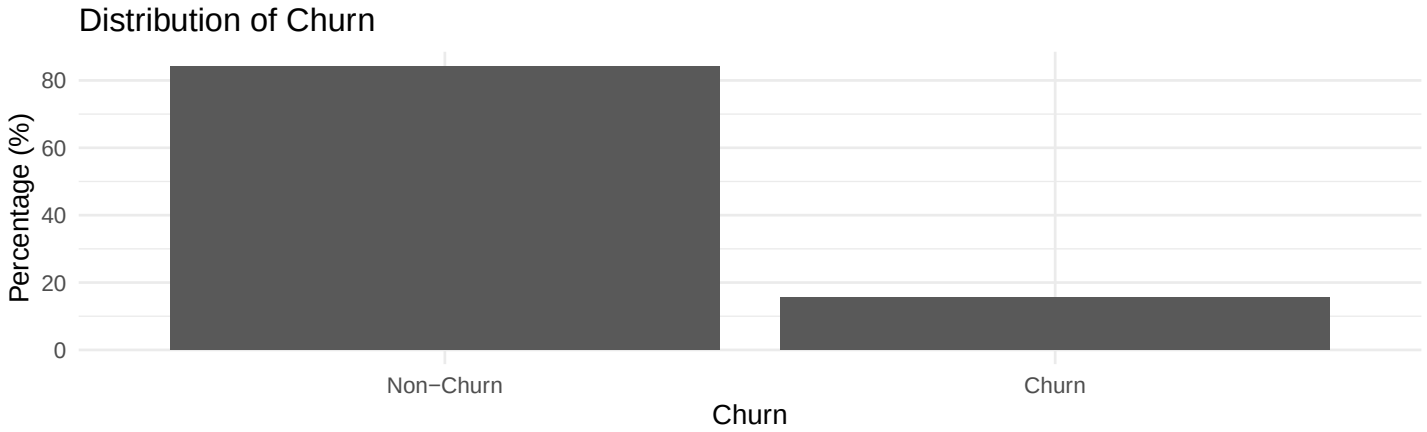
For each one-unit increase in Subscription Length, the expected log of the Customer Value decreases by 0.01327.The coefficient's p-value 0.1229 does not indicate statistical significance. For a customer switching from Low Charge Amount to High Charge Amount, the expected log of the Customer Value increases by 0.8957. The coefficient's p-value 4.69e-08 indicates statistical significance. Every additional second of use is associated with a 0.0003162 increase in the log of the response variable. The coefficient's small p-value indicates statistical significance. Each additional SMS sent is associated with a 0.01086 increase in the log of the response variable. The coefficient's small p-value indicates statistical significance. Being on this contractual plan, as opposed to the pay-as-you-go plan, is associated with a 1.928 increase in the log of the response variable. The coefficient's p-value 0.0018 indicates statistical significance. The interaction between High Charge Amount and being on this contractual plan is associated with a decrease of 2.108 in the log of the response variable. The p-value of 0.0020 indicates statistical significance.

Model Fit

The final model yields an adjusted R^2 of 0.2156, meaning that approximately 21.56% of the variation in the dependent variable is explained by the model. This seem modest.

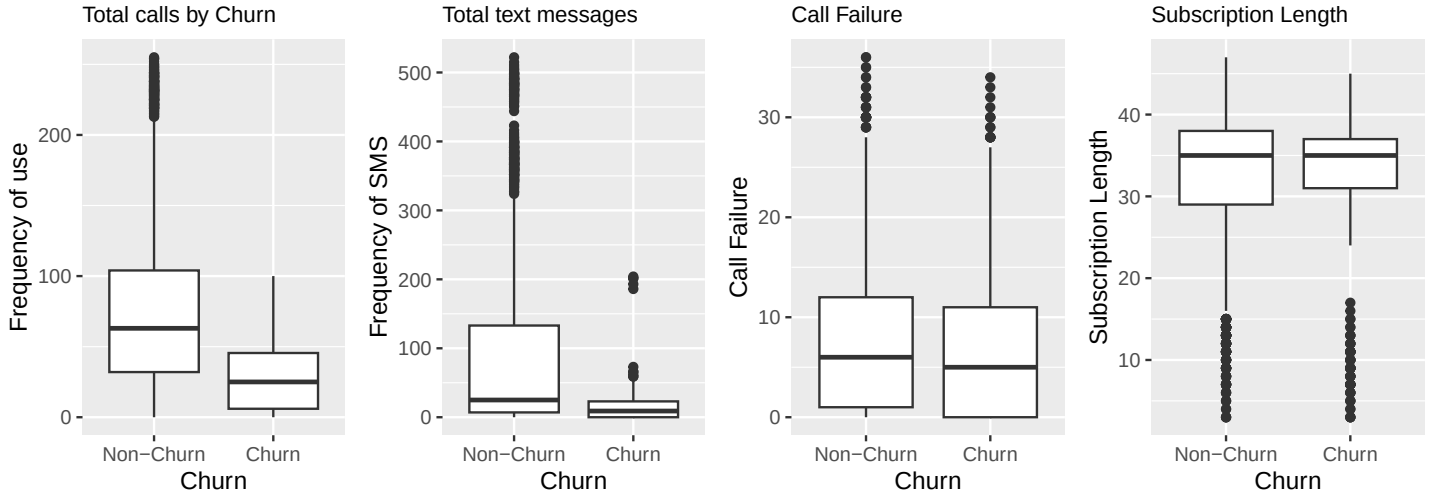
4.2 *How can we forecast customer churn effectively?*

4.2.1 Exploratory Data Analysis



In our dataset, we find that 495 customers of this telecom company chured, which consists 15.71%. On the other side, we have 2,655 customers decided to stay. In the following plots, we can observe that for predicting how likely a customer will churn, the more a person uses the service in terms of both the number of calls and SMS, the less

likely they are to become a churned customer. Then, we also want to add factors, like number of call failures, subscription length, charge amount, and so on in our proposed model to provide more accurate prediction.

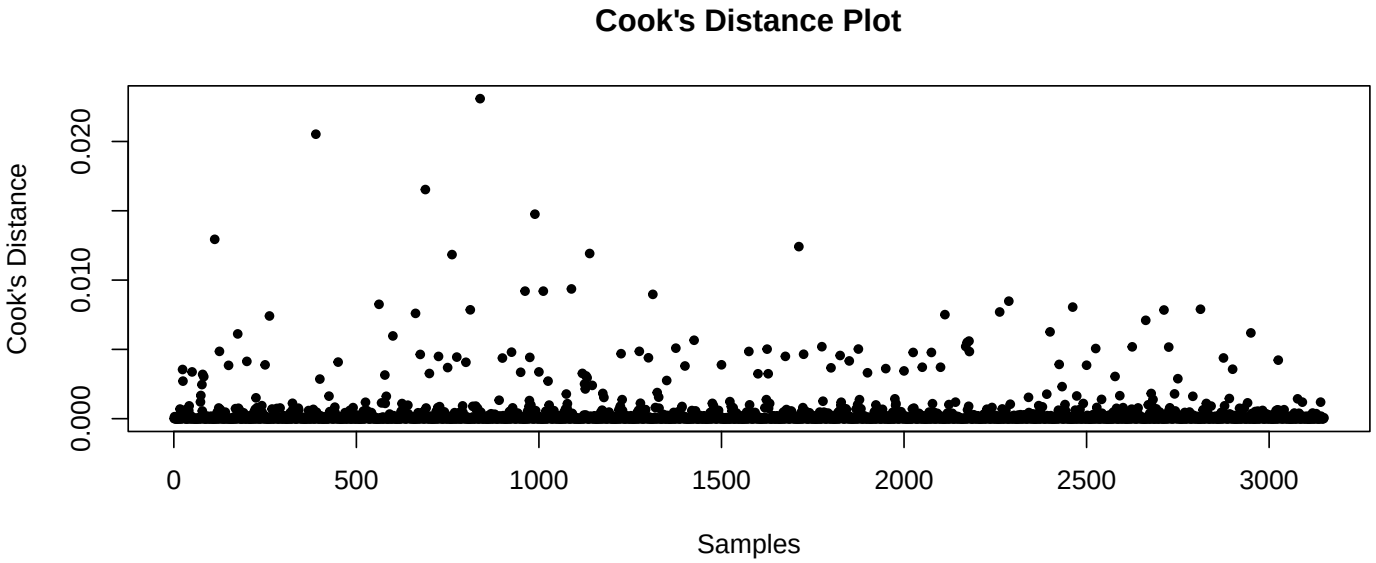


4.2.2 Model

Based on our EDA findings, we employ factors, including number of call failures, subscription length, charge amount, second of use, frequency of use, frequency of SMS, tariff plan, and the call failure complaints in our proposed model to provide more accurate prediction. So we first build a baseline model. After fitting this model, we can get it's performance and will be able to further conduct steps, like feature engineering, removing influential points, and so on.

Influential points

One concern about the model traning is that outliers can have a large impact on the model. So we attempt to identify influential points based on Cook's distance. We consider 0.01 can be a reasonable threshold value from the visualization of all samples' Cook's distance, if the Cook's distance of a data point is larger than the threshold value, we should mark it as a influential point. In our dataset, 10 samples are recognized as influential points, since they only constitute 0.3% of the whole data, we can safely drop them all.



4.2.3 Model Assessment

Model Summary

After get the model with highest accuracy and kappa from the cross-validation, we find that traiff plan should be excluded in the model. The regression result is shown as below:

Variable	Coefficient	Standard Error
Call Failure	0.112***	(0.016)
Complains	5.369***	(0.569)
Subscription Length	-0.028***	(0.010)
Charge Amount	-0.309***	(0.093)
Frequency of use	-0.027***	(0.004)
Frequency of SMS	-0.021***	(0.004)
Status	1.337***	(0.196)
Call Failure * Complains	-0.119***	(0.037)
Constant	-0.732**	(0.304)
Observations	3,142	
Log Likelihood	-685.529	
Akaike Inf. Crit.	1,389.059	
Accuracy	0.8571	
95% CI	(0.8444, 0.8692)	
Kappa	0.5801	

Note: p<0.1; p<0.05; p<0.01

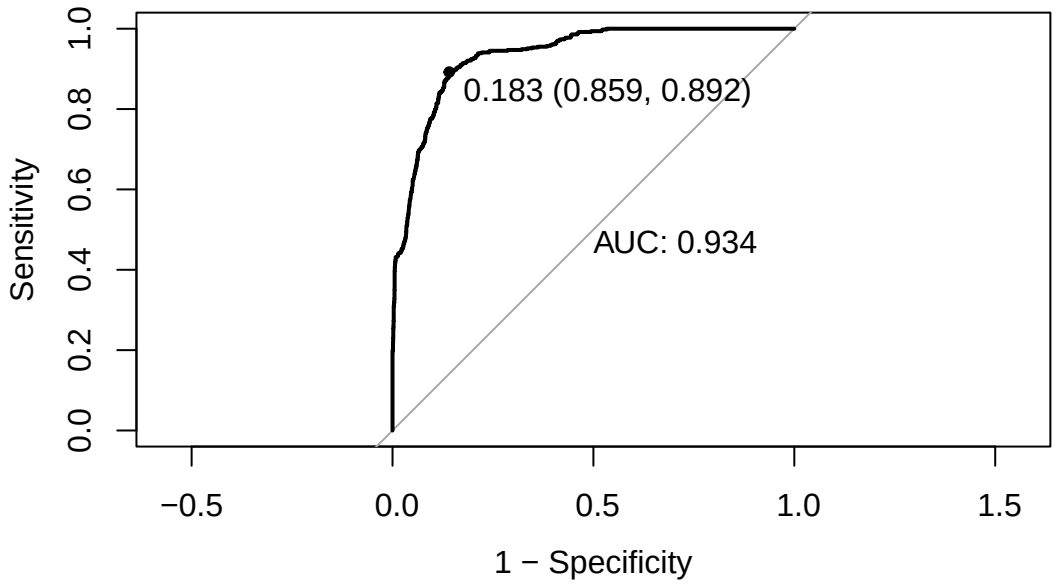
Model Interpretation

Here are some important findings about coefficents and odd ratios: - Customers who register a **complaint** about their telecommunication service are approximately **214 times more likely** to churn in the future.

- Each additional **call failure** corresponds to a **11.9% increase** in the probability of customer churn.
- An important discovery is the concept of **"User Stickiness"**. The longer a customer stays with this company, the less likely they are to churn. For each additional month of subscription, there is a **3% reduction** in the likelihood of churn.
- Increased **frequency of service use**, whether for calls or SMS, is associated with a decreased likelihood of churn.
- A one-unit increase in the **charge level** results in an approximately **27% decrease** in the probability of churn.
- **Inactive users** are approximately 3.8 times more likely to churn.

Model Fit

For testing our model performance, we focus on accuracy and kappa. The accuracy is 85.71% and it's 95% confident interval is (84.44%, 86.92%). Our model's kappa is 0.5801. From the out-of-sample perspective, we can split our dataset to a training set of 80% data and testing set of 20% data. We train the model in training set and test in the testing set. The out-of-sample accuracy is 86.31% and the kappa is 0.5867. hThis result reveals that our model has a strong predictive power and can effectively forecast the customer churn.
Here we plot the ROC-AUC to visual our model performance. Finding the best threshold at 0.178, we find that the AUC is 0.933.



In summary, we build a logistics model that can effectively predict whether a customer will churn. We solve the multi-colinearity with vifs and outliers using Cook's distances. A variable selection is also conducted with cross-validation. So we finally get a predictve model with 85.71% accuracy, 0.5801 kappa and 0.933 AUC. In our proposed model, we explain odd ratios and prove the user stickiness. The longer our customer subscripe our telecom service and the more they use, the less likely they will leave us.

5 Conclusion

5.1 Key Insights

In our examination of Customer Value, it has been ascertained that variables such as high charge amounts, extensive service usage, and engagement in contractual plans are pivotal in augmenting the value derived from telecom customers.

Conversely, our analysis in Churn Prediction has pinpointed critical indicators that may forecast a customer's propensity to discontinue services. These indicators encompass factors such as inactivity, the incidence of complaints, and the frequency of call failures. Notably, these factors are demonstrative of the potential for customer attrition and underscore the importance of proactive measures to ameliorate service quality and retain clientele.

5.2 Limitations of the Study

It is imperative to acknowledge the limitations inherent within our study. The skewed nature of the data set forth challenges, and the specificity of the industry context may restrict the generalizability of our conclusions to other sectors or geographic regions. Moreover, our insights are contingent upon the scope of the dataset provided, which may not encompass all possible influential factors due to limitations in data accessibility.

5.3 Prospects for Future Research

As we project into the future, several avenues for extended research emerge. An enriched analysis incorporating external variables not included in the present study may yield more refined predictive models. Furthermore, it would be advantageous to undertake a longitudinal study to observe and interpret the evolution of customer behavior over extended periods. Such an approach would afford a more nuanced and dynamic understanding of customer interactions and their long-term implications on the telecommunications industry.

6 References

UCI Machine Learning Repository. (2020). Iranian Churn Dataset (No. 563). https://doi.org/10.24432/C5JW3Z

7 Appendix

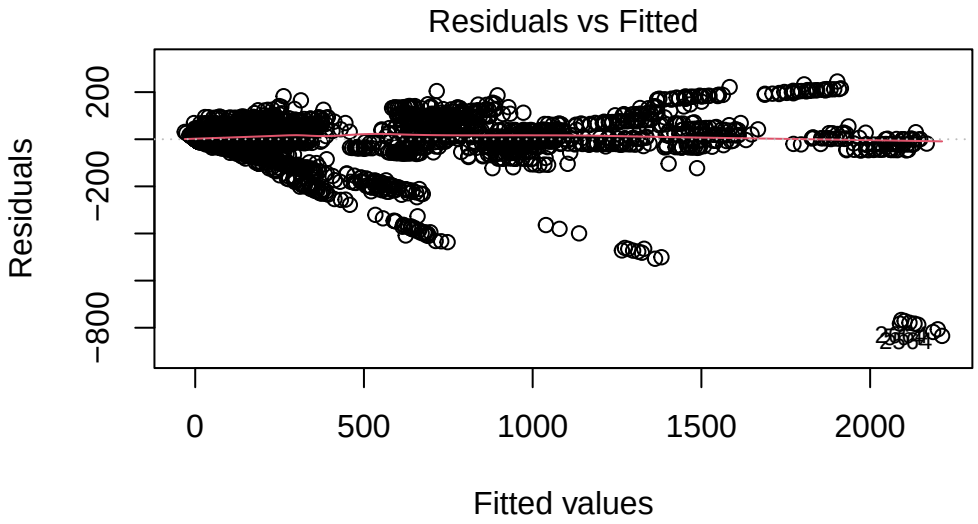
7.1 Variables

The 14 variables in our dataset and their characteristics are detailed in the table below.

Variable	**Description**	**Type**	
Call Failure	Number of call failures	Numeric	Discrete
Complains	(0: No complaint, 1:complaint)	Categorical	Nominal
Subscription Length	Total months of subscription	Numeric	Discrete
Charge Amount	(0: lowest amount, 9: highest amount)	Categorical	Ordinal
Seconds of Use	Total seconds of calls	Numeric	Discrete
Frequency of use	Total number of calls	Numeric	Discrete
Frequency of SMS	Total number of text messages	Numeric	Discrete
Distinct Called Numbers	Total number of distinct phone calls	Numeric	Discrete
Age Group	(1: younger age, 5: older age)	Categorical	Ordinal
Tariff Plan	(1: Pay as you go, 2: contractual)	Categorical	Nominal
Status	(1: active, 2: non-active)	Categorical	Nominal
Age		Numeric	Discrete
Customer Value	The calculated value of customer	Numeric	Continuous
Churn	(1: churn, 0:non-churn)	Categorical	Nominal

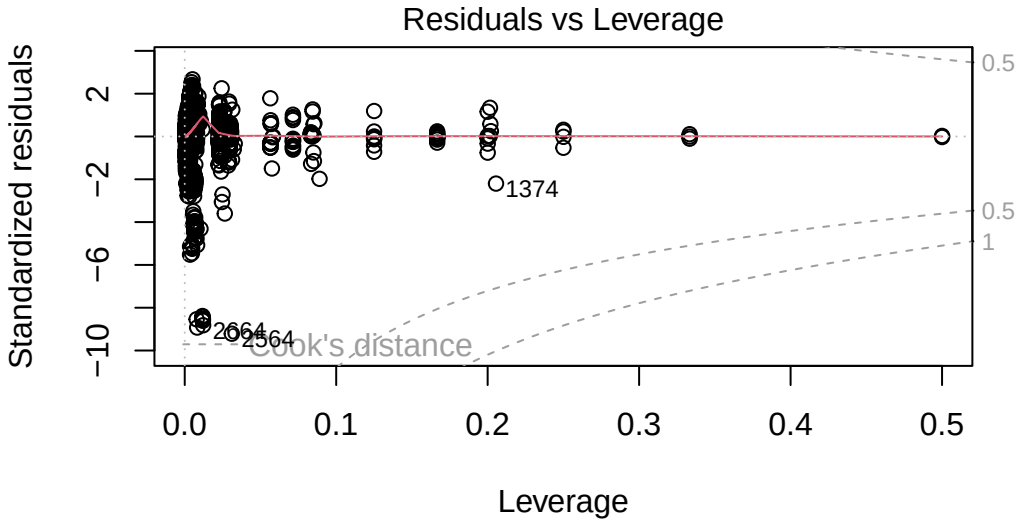
7.2 Plots

Residual Plot for Linear Regression Model of Customer Value



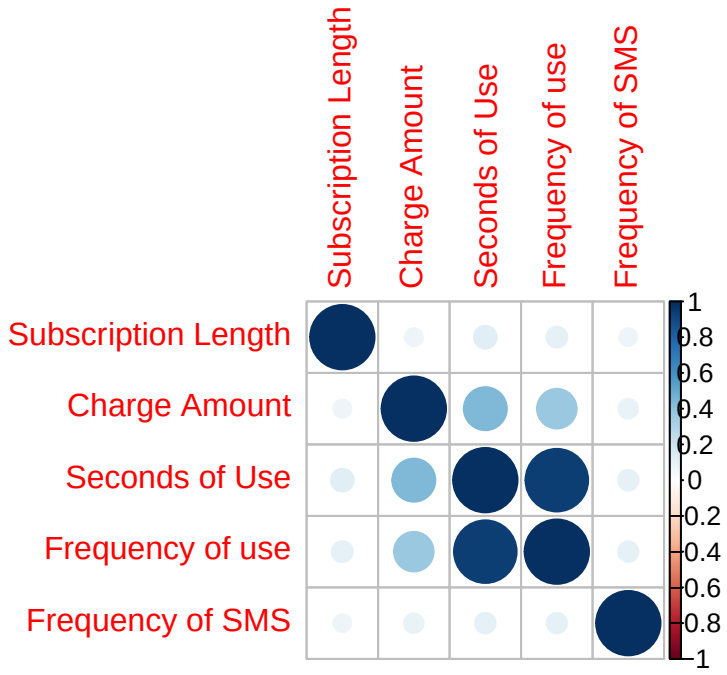
n('Customer Value' ~ 'Subscription Length' + ChargeAmountFactor + 'Secc

Cook's distance for Linear Regression Model of Customer Value

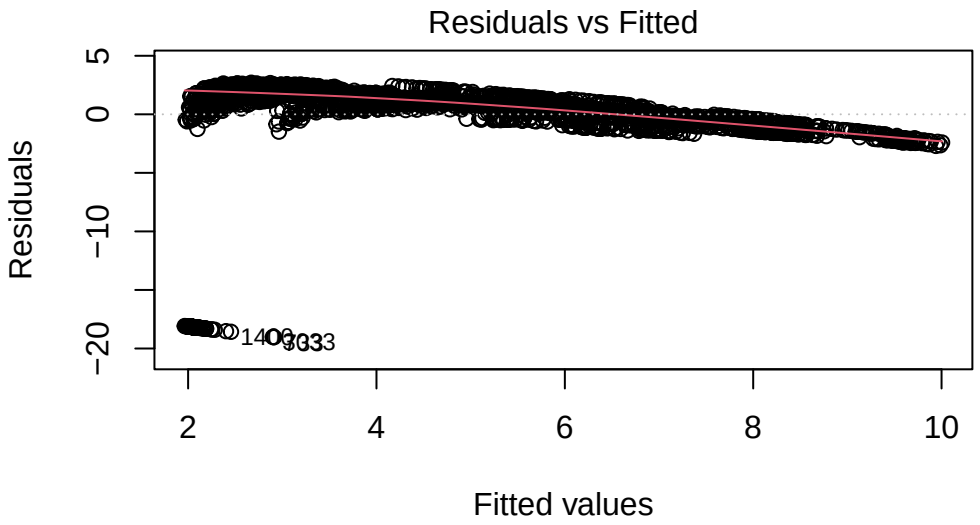


n('Customer Value' ~ 'Subscription Length' + ChargeAmountFactor + 'Secc

Correlation Matrix and VIF Analysis for Linear Regression Model of Customer Value



Residual Plot for Linear Regression Model of Log of Customer Value



n(log(‘Customer Value‘ + 1e−07) ~ ‘Subscription Length‘ + ChargeAmountE