Data Science - Assessed Coursework 2

Statistical Analysis Report on Iranian Churn Dataset

CID: 01891103

Abstract

This report investigates customer churn patterns within an Iranian telecom dataset available on the UCI machine learning repository. The main aim is to model and predict the reasons for customer churning behaviour using various binary classification methodologies. A detailed exploratory data analysis (EDA) identified several skewed variables across varying scales and high correlations between covariates. Consequently, a square root transformation and standardisation are conducted before the modelling procedure. In this report, we implemented and compared the performance of the Stepwise Selected models (AIC & BIC), Grouped LASSO (`gglasso`) and Support Vector Machine (SVM) models. The SVM model gives the best AUC while the BIC model predicts to gives the best accuracy on the test set. Further suggestions and concerns are provided at the end of this report.

1 Introduction

The Iranian Telecom Company concerns a long-standing issue of managing customer churning, i.e. switching to other telecom providers. To better maintain customer loyalty and implement effective customer retention strategies, it would be crucial to know why customers churn and predict who will possibly churn based on the data collected by the company from the customers.

Therefore, to help the Iranian telecom company to better predict customer behaviour hence increasing their profit, this report gives a detailed analysis of the churning behaviour dataset. The data is provided and maintained by (Jafari-Marandi 2020) and is open-source available on the UCI machine learning repository.

2 Exploratory Data Anlaysis

The dataset contains 3150 observations (customers) with 14 columns, with no missing values in any cell. The columns consist of 1 column for the outcome binary indicator on the churning behaviour of the customers, 1 column indicating the Customer Value which is not an attribute to the modelling process according to (Jafari-Marandi 2020), and 12 feature/attribute columns describing the customer information including. The attributes are summarised in the Table 1

Our dataset is split into 80-20 train-test split subsets, during the Exploratory Data Analysis (abbreviated as EDA later) sections, the **whole** dataset is analysed to detect the full patterns and provide a comprehensive overview of the dataset. Later in the modelling section, only the **training** set is used to select hyperparameters and train the models, where the **test** set is used only for examining the performance of the trained model and to compare their performance.

Variable Name	Type	Description
Call Failure	continuous numeric	number of call failures
Complains	binary	0 = No complaint, 1 = Complaint
Subscription Length	continuous numeric	total months of subscription
Charge Amount	ordinal numeric	0 lowest amount, 9 highest amount
Seconds of Use	continuous numeric	total seconds of calls
Frequency of use	continuous numeric	total number of calls
Frequency of SMS	continuous numeric	total number of text messages
Distinct Called Numbers	continuous numeric	total number of distinct phone calls
Age Group	ordinal attribute	¡15, 15-30, 30-45, 45-60, 60-75
Tariff Plan	binary	1 = Pay as you go, 2 = contractual
Status	binary	1 = active, 2= non-active
Churn	binary Class label	1 = churn, 0 = non-churn

Table 1: Summary of Outcome and Feature variables in the Dataset

2.1 Univariate Analysis - distributions of variables

From the univariate analysis plots, we notice that most continuous variables have high positive skewness, except the subscription length. This indicates that we may need to investigate further on the empirical logit plot later to decide if a transformation or scaling is needed.

Also, the distribution of the continuous variables has very different scales, e.g., from 0-40 (call failures) to 0-20000 (Seconds of Use). This requires attention and consideration of standardization procedure before implementing some models which are sensitive to scales such as LASSO models.

Finally, the bar charts of the binary variables, including the outcome variables, all show a strong imbalance between the two levels. This may raise concerns regarding the predictive power of any model built on this dataset.

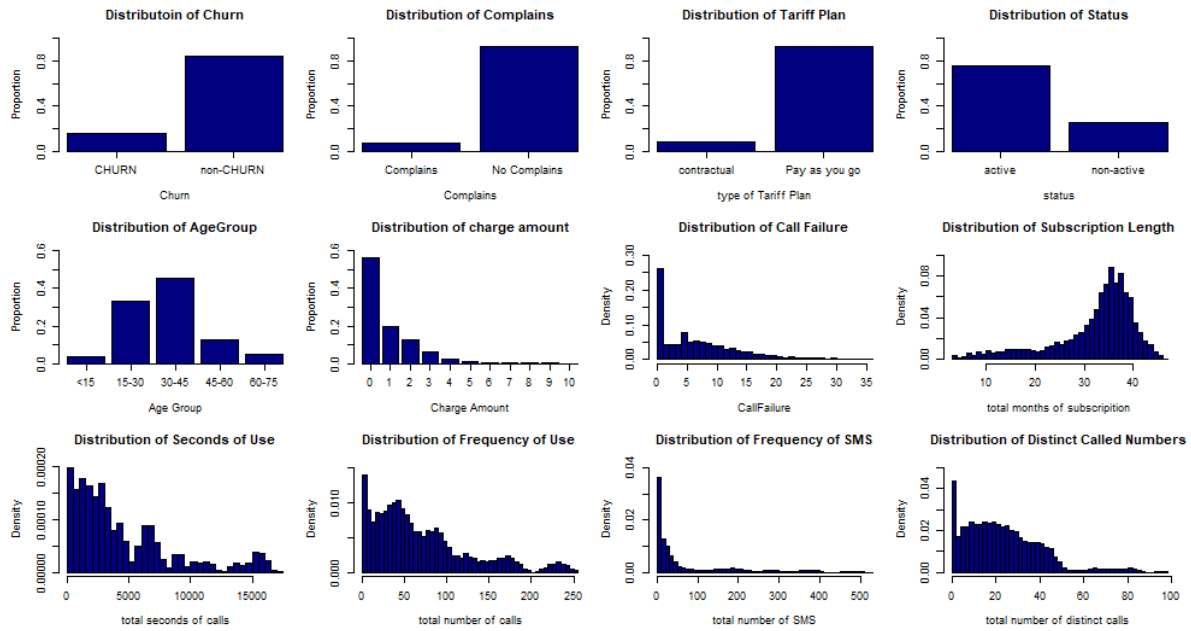

Figure 1: Univariate Plot of the variables.

2.2 Bivariate Analysis - correlations between variables

Now instead of investigating single variables, we explore the relationships between two variables. The two important bivariate analysis topics are 1) the relationship between outcomes and the covariates, and 2) the correlations in-between the covariates.

Below is a very informative plot demonstrating the most relevant information in all continuous variables and the outcome Churn.

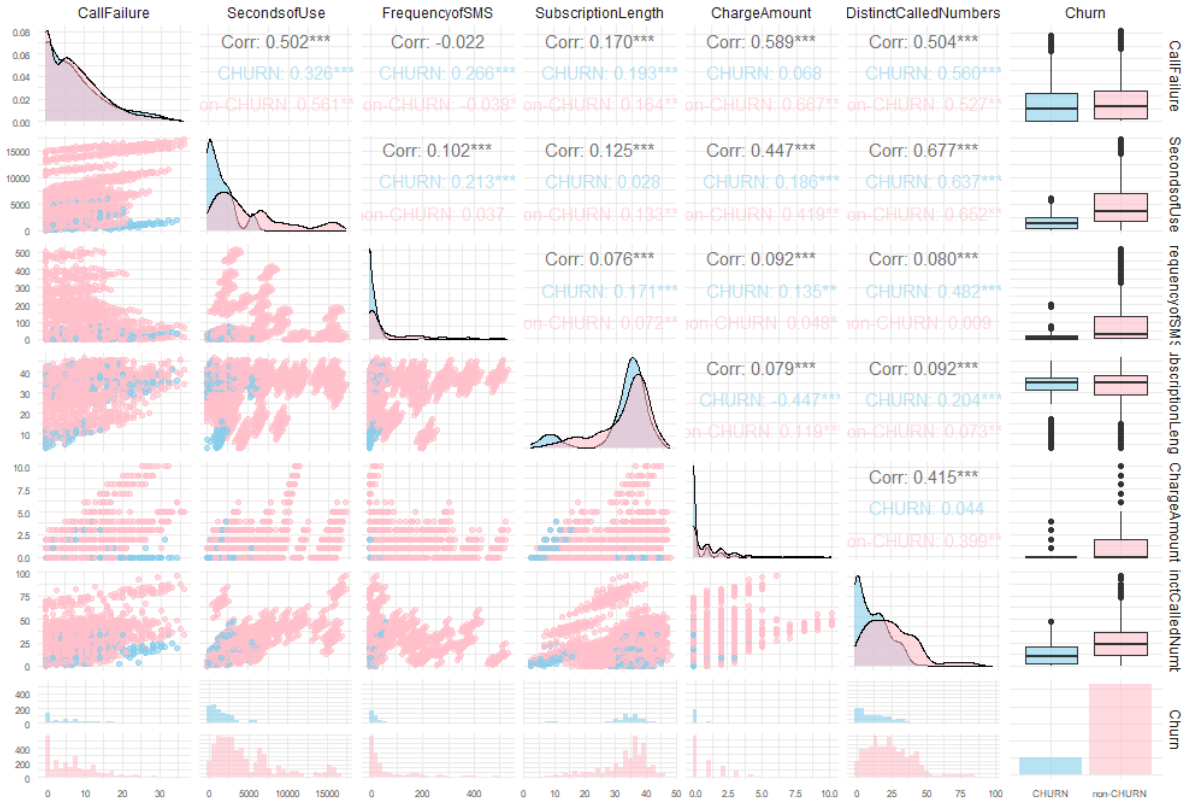

Figure 2: Paris plot of all continuous variables and the outcome Churn

Most features show a different distribution with respect to different customer churning behaviour. For example, in the case of total frequency/seconds of calls, observe that customers who churned (blue) demonstrate a clearly lower usage of calling services compared to non-churn customers (pink).

Also, notice how the company indeed received more complaints in the churned group, who also use the the "pay as you go" plan instead of subscriptions. This agrees with common sense as it indicates a lower brand loyalty.

Multicollinearity

Notice also that the multicollinearity problem exists in our dataset. Figure 3 gives the pairwise correlation information of all continuous variables in the dataset. All pairs of covariates exhibit a positively correlated relationship to some extent.

Specifically, it is very clear that the frequency of use and the second of use gives a correlation almost equal to one. This is not a surprise as they basically predict the same thing. If both variables are included in the model, then we may need to pay close attention to the independence assumption of the (generalised) linear model which is very likely violated.

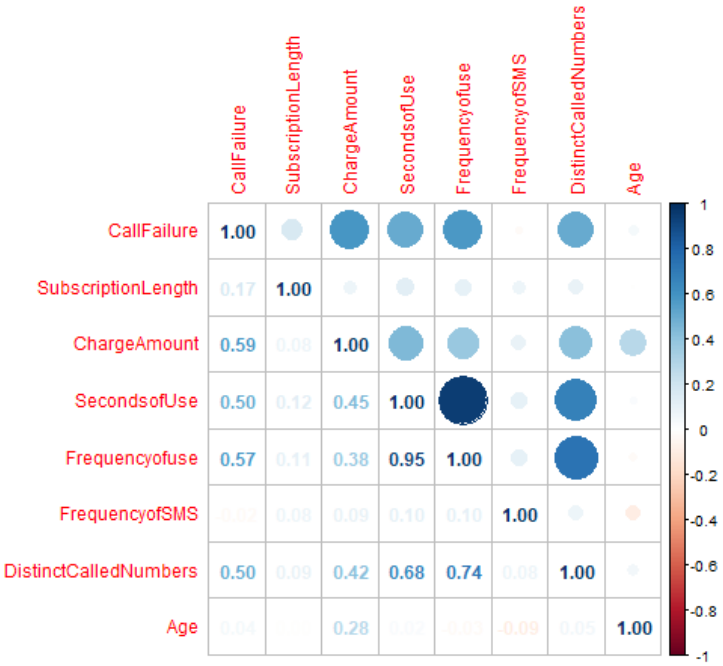

Figure 3: Multicollinearity heatmap of all continuous covariates

Additionally, the subscription length is uncorrelated with most other (continuous) variables.

2.3 Transformation - linearity assumption for GLMs

The following initial empirical logit plots shows that most of the relationships between the link and the variables are non-linear, e.g., months of subscription, seconds/frequency of calls/SMSs etc. This violates the linearity assumption and hence needed to be transformed before fitting into the GLM model.

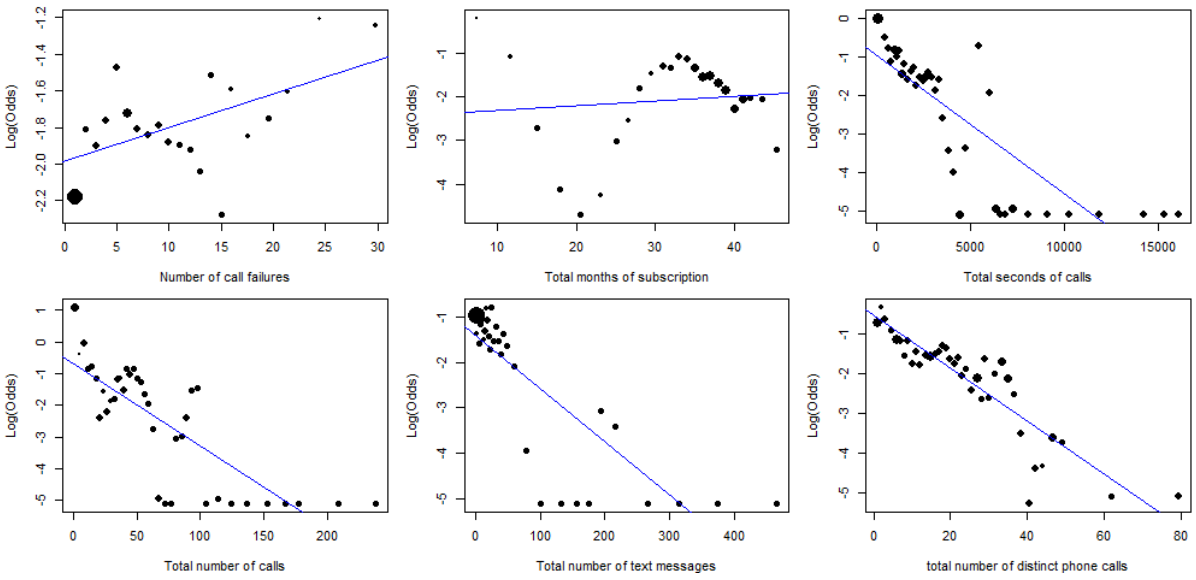

Figure 4: Initial Empirical Logit plots

Notice that the months of subscription variable clearly shows a cubic relationship with the links,

(logit($\mathbb{P}$(Churn)), which indicates that a cubic polynomial term might be needed for a logistic regression model.

Therefore, we tried several different transformation methods, including square-root transformation, log transformation and the inverse transformation, and the transformed empirical logit plot, Figure 5 indicated the relationship between the square-root transformed variable and the logit link is the most linear with the least violation of the linearity assumption.

In conclusion, we flagged that the `Seconds of Use`, `Frequency of use`, and `Frequency of SMS` variables need to be square-root transformed before logistic modelling.

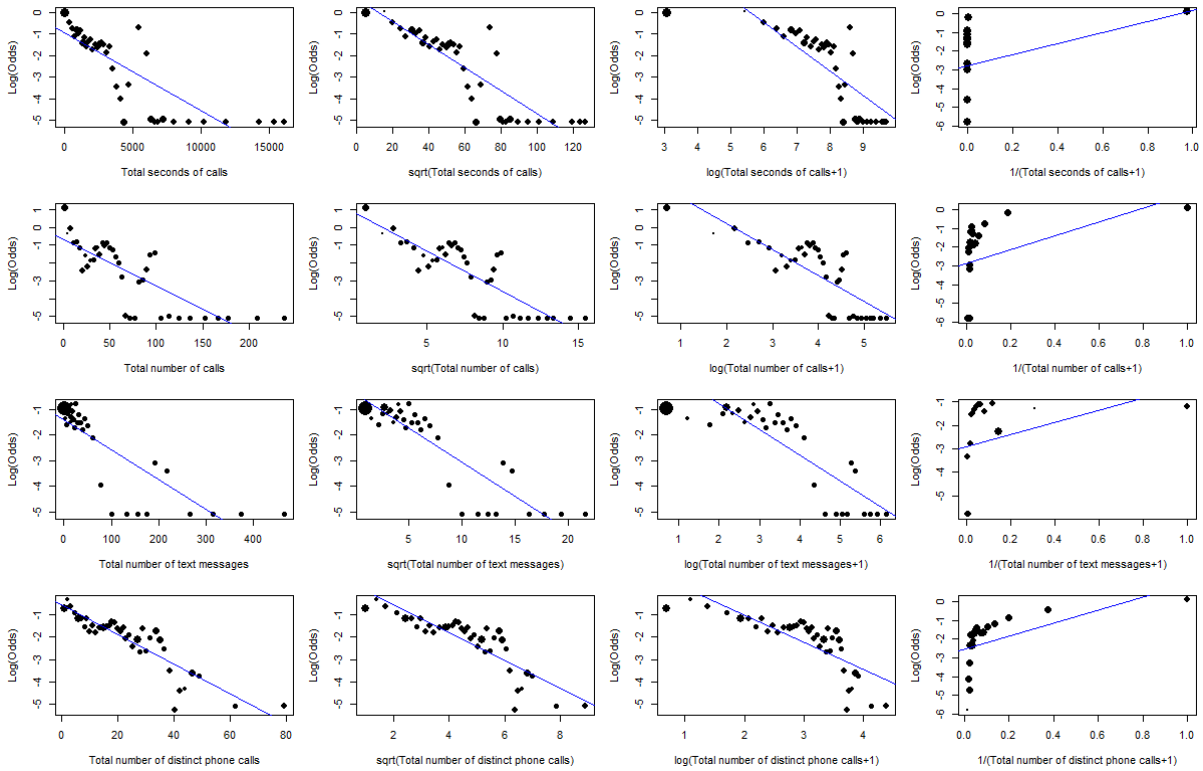

Figure 5: Empirical logit plot for continuous variables

Notice that standardisation is needed in some models (discussed later in Section 3), we also plot the empirical logit plot after transformation and then standardisation in Figure 6.

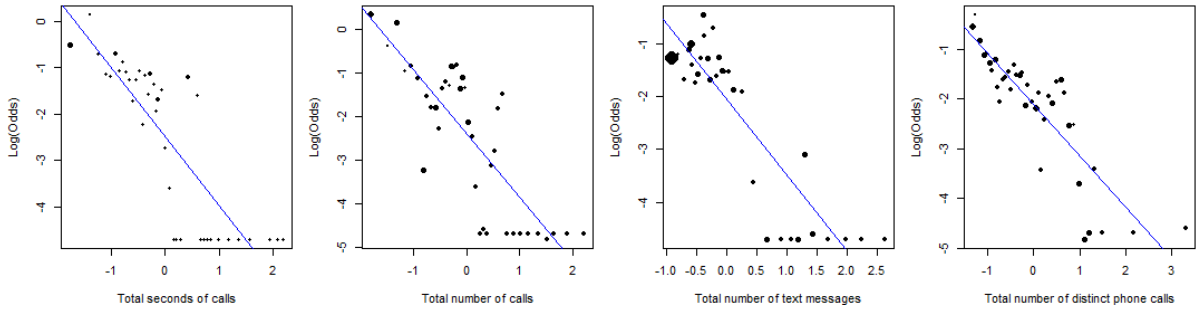

Figure 6: Empirical Logit plot of variables after transformation and standardisation (left 1-3 figures), or standardisation only (rightmost)

2.4 Train-test Split Check

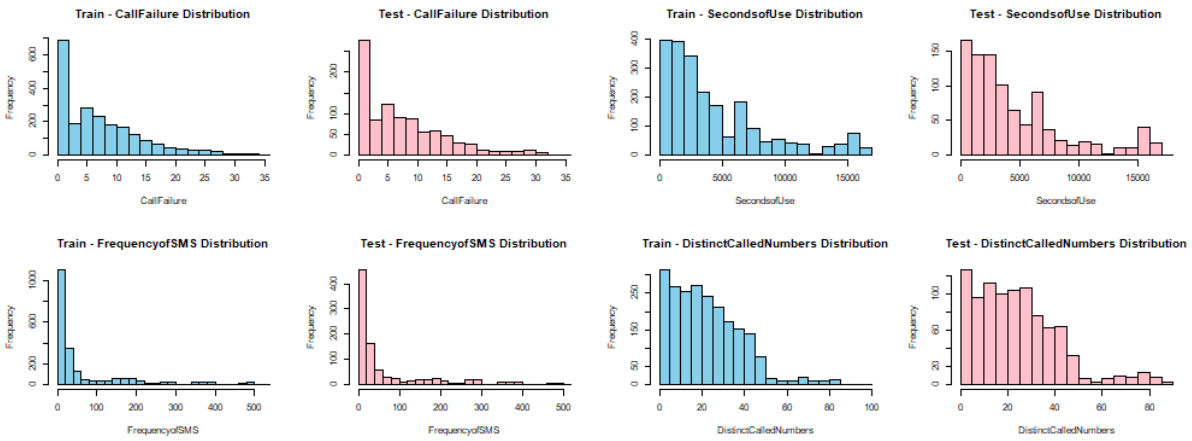

Figure 7: Checking test dataset mimics the training dataset for discrete variables

One more thing we need to check in the EDA section is that the test set mimics the patterns and distributions of the training dataset. This is to ensure that our model performance check on the test set is valid, representative, and generalisable. Figure 7, and Figure 8 shows that the univariate distribution of the variables are correctly represented in the test set.

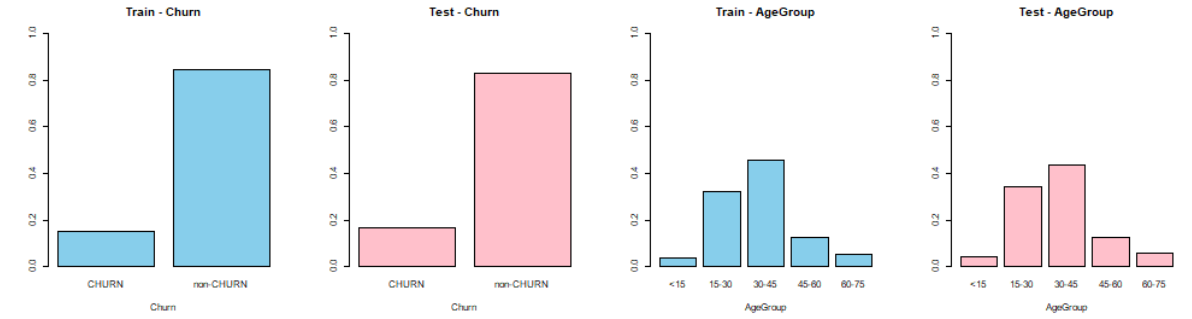

Figure 8: Checking test dataset mimics the training dataset for discrete variables

3 Model Fitting

After careful consideration, in this report we will implement and test for 4 models, they are:

1. AIC stepwise selected model
2. BIC stepwise selected model
3. LASSO penalised regression model
4. Support Vector Machine (SVM) model with kernel trick

Notice that, in this section, we use the training set only. As described in the EDA Section 2, 3 variables are pre-transformed using square-root transformation for **model 1, 2, and 3**, and then all continuous variables are then standardised to mean 0 with standard deviation 1 for **model 3**. The **model 4** use the dataset with standardisation procedure without transformation.

3.1 Initial Full model

With adjusted for all transformations, polynomial terms, and including all available features in our model, we now obtain the initial full logistic regression model as below:

$$
\begin{aligned}
\text{logit}(\mathbb{P}(\text{Churn})) = \beta_0 &+ \beta_1 \texttt{CallFailure} + \beta_2 \mathbf{1}_{\texttt{Complains=yes}} + \beta_3 \texttt{SubscriptionLength} \\
&+ \beta_4 \texttt{SubscriptionLength}^2 + \beta_5 \texttt{SubscriptionLength}^3 + \beta_6 \texttt{ChargeAmount} \\
&+ \beta_7 \mathbf{1}_{\texttt{Age in 15-30}} + \beta_8 \mathbf{1}_{\texttt{Age in 30-45}} + \beta_9 \mathbf{1}_{\texttt{Age in 45-60}} + \beta_{10} \mathbf{1}_{\texttt{Age in 60-75}} \\
&+ \beta_{11} \mathbf{1}_{\texttt{TariffPlan=contractual}} + \beta_{12} \mathbf{1}_{\texttt{Status=non-active}} \\
&+ \beta_{13} \texttt{Distinct Called Numbers} + \beta_{14} \sqrt{\texttt{Second of Use}} \\
&+ \beta_{15} \sqrt{\texttt{Frequency of use}} + \beta_{16} \sqrt{\texttt{Frequency of SMS}}
\end{aligned}
$$

3.2 Stepwise Selection

Stepwise selection involves iterative processes to either add or drop variables based on information criteria like AIC or BIC. Our stepwise selection alternates between forward and backward direction (`direction = both`) to move toward the model which improves the criterion the most, and stopped until AIC/BIC could not be improved by adding to dropping variables any further.

3.3 Grouped Lasso Method

Just like Lasso variable selection method, one more step that the `gglasso` method allows us to select multiple parameters together, i.e. either select the group of parameters together or send all parameters in the group to exactly 0. This allows us to perform variable selection via L1 norm penalised regression modelling for original categorical variables.

As a convention, we use the λ (penalise parameter) which gives Mean Square Error (MSE) within 1 standard deviation of the minimum MSE (right dotted line Figure **??**). This is to prevent overfitting and give a more parsimonious model within a reasonable range.

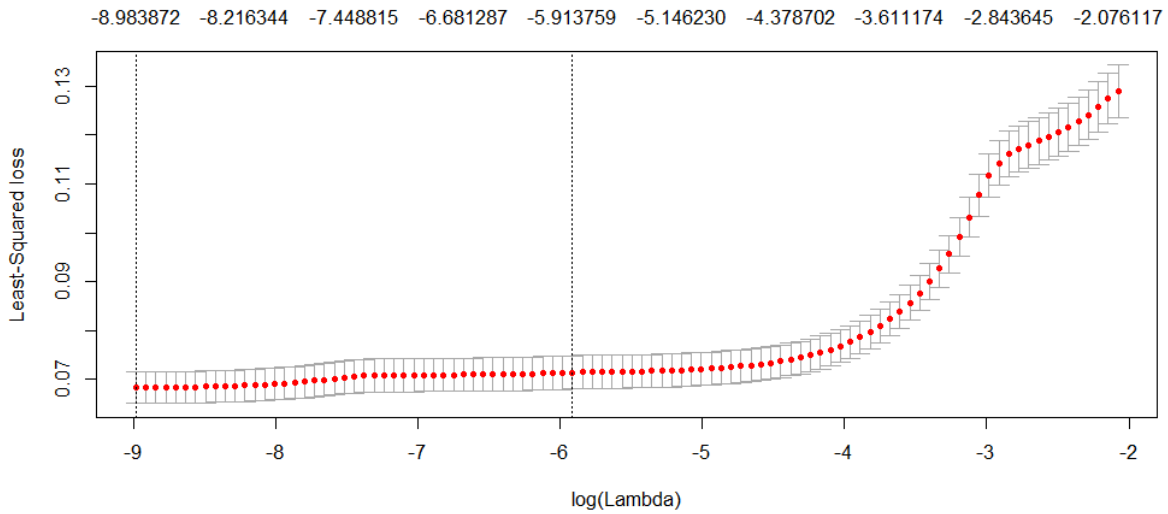

Figure 9: Cross-validation result for Grouped Lasso method.

3.4 Support vector machine

The strengths of the Support Vector Machine (SVM) include its effectiveness in high-dimensional spaces and versatility in handling various types of data and relationships (e.g., non-linear) by the kernel tricks. Hence it is particularly suitable for the binary classification task in our case, as our data has several non-linearity problems as discussed in the transformation section.

SVM operates by finding the hyperplane that optimally separates different classes in the feature space. It maximizes the margin between the nearest data points of any class (known as support vectors).

Notice also, in this modelling process, we do not transform the variables, but only standardise the continuous feature variables since SVM is also sensitive to scale due to the notion of maximum margin.

3.5 Vairable Selection Summary

After implemented the above models, the summary of variable selection results are as below:

Variable	**AIC**	**BIC**	**gglasso**	**SVM**
Call Failure	✓	✓	✓	✓
Complains	✓	✓	✓	✓
I(Subscription Length)^3	✓	✓		✓
Charge Amount	✓	✓	✓	✓
Seconds of Use			✓	✓
Frequency of use	✓	✓	✓	✓
Frequency of SMS	✓	✓	✓	✓
Distinct Called Numbers		✓		✓
Age Group	✓			✓
Tariff Plan				✓
Status	✓	✓	✓	✓

Table 2: Summary of Outcome and Feature variables in the Dataset

We noticed that the `gglasso` select the least number of variables, and the SVM model does not perform variable selection hence all variables are selected.

Also, some variables are selected by all variable selection process, they are: Call Failure, Complains, Charge Amount, Frequency of use, Frequency of SMS and Status.

4 Model Evaluation

4.1 Model Performance

Now, in this section, we use the test set only.

Our primary goal is to examine the performance of the models. Hence to do so, we follow the following steps:

1. predict the fitted response, i.e. the probability of Churn, by using the trained models 1 to 4 based on the feature variables in the test set
2. assign the customer to Churn group if the probability of churning is ¿ 0.5, and to non-Churn group is ¡=0.5
3. produce the confusion matrix of the models and calculate the accuracy of the prediction.
4. plot the ROC curve of the binary classification model, and calculate the corresponding AUC.

Given the 0.5 threshold, the 4 confusion matrix is given by:

AIC model:

		True Label	
		non-Churn	Churn
Prediction	non-Churn	770	69
	Churn	16	90

Accuracy: 91.01% Sensitivity: 97.96% Specificity: 56.60%

BIC model:

		True Label	
		non-Churn	Churn
Prediction	non-Churn	776	74
	Churn	10	85

Accuracy: 91.11% Sensitivity: 98.73% Specificity: 53.46%

gglasso model:

		True Label	
		non-Churn	Churn
Prediction	non-Churn	776	104
	Churn	10	55

Accuracy: 87.94% Sensitivity: 98.73% Specificity: 34.59%

SVM model:

		True Label	
		non-Churn	Churn
Prediction	non-Churn	782	104
	Churn	4	55

Accuracy: 88.57%, Sensitivity: 99.49%, Specificity: 34.59%

summary on confusion matrices:

The highest accuracy is given by the BIC model, 91.11% accuracy.

The `gglasso` model and the SVM model gives a higher sensitivity but lower specificity compared to the stepwise selected model using a 0.5 probability threshold. However, considering the high cost of NOT applying any retention strategy to the churning customers, a higher specificity is normally considered more important than a minor improvement in sensitivity.

Hence, in the analysis of model performance metrics, the BIC model gives the best performance due to its high accuracy and high specificity.

ROC and AUC analysis

In addition, we may want to the ROC plots of the 4 models are given in Figure 10. Clearly, the SVM model gives the best roc with the highest area-under-curve (AUC) value.

We notice that, actually in the specificity Vs sensitivity analysis, the model performs both higher sensitivity and higher specificity is the SVM model. Improving both values can be done by moving the classification threshold.

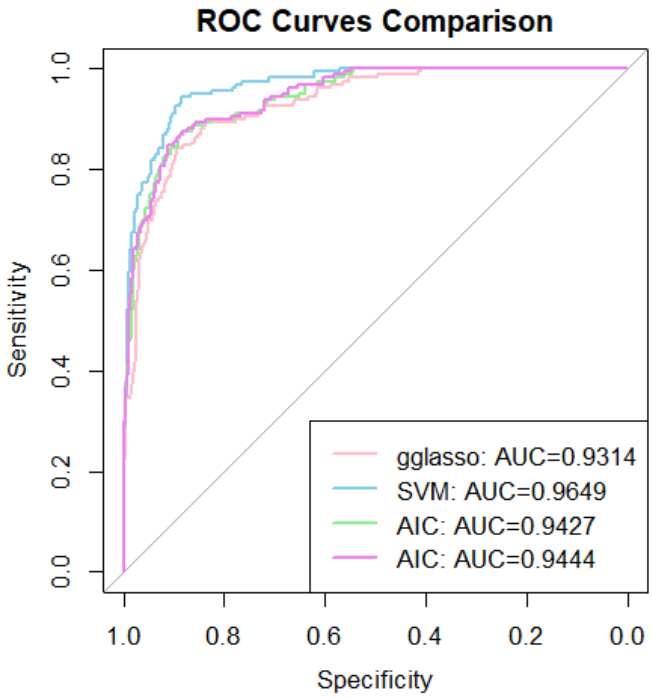

Figure 10: The ROC plots for all 4 models

4.2 Model Diagnostics

To confirm that the BIC model is indeed a good fit for the dataset, we also need to make sure that the model meet the assumptions well. Below is the diagnostics plots for the BIC model, and we may conclude from the following plots that the model assumptions are properly met, since:

- Cook's distances are all very small (< 0.05).
- The influence plot seems to have no outliers with high leverages.
- all points fall in the confidence interval of the half-normal plot and it shows a reasonable linear relationship.

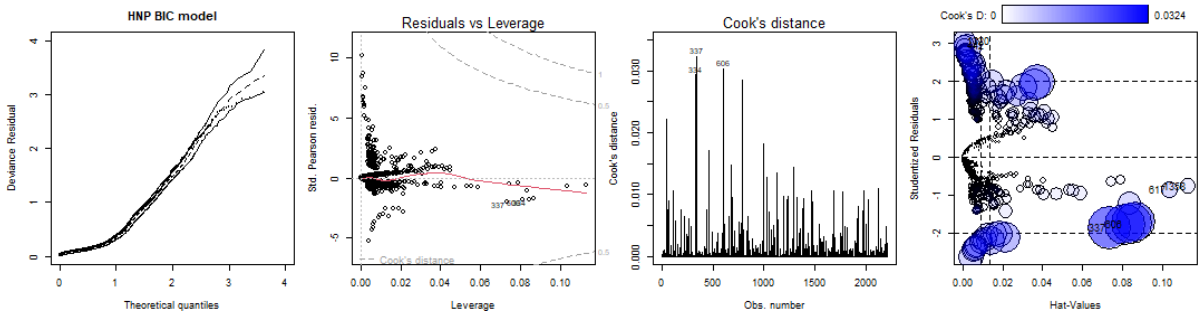

Figure 11: Model Diagnostics plots for the Stepwise BIC model.

5 Conclusion and interpretations

Both the SVM model and the BIC model give high performance, however the SVM model gives a better ROC with different decision thresholds. This is not part of our analysis in this report.

Therefore, based on all analyses before, we chose the BIC model as our final model, below is

the final model formula and the coefficients.

$$\begin{aligned}\text{logit}(\mathbb{P}(\text{Churn})) = \beta_0 &+ \beta_1 \texttt{CallFailure} + \beta_2 \mathbf{1}_{\texttt{Complains=yes}} + \beta_3 \texttt{SubscriptionLength} \\ &+ \beta_4 \texttt{SubscriptionLength}^2 + \beta_5 \texttt{SubscriptionLength}^3 + \beta_6 \texttt{ChargeAmount} \\ &+ \beta_7 \mathbf{1}_{\texttt{Status=non-active}} + \beta_8 \sqrt{\texttt{Frequency of use}} + \beta_9 \sqrt{\texttt{Frequency of SMS}}\end{aligned}$$

Table 3: Model Estimates

Variable	**Estimate**
(Intercept)	0.02549
Call Failure	1.12131
Complains: No Complains	-3.91967
poly(Subscription Length, 3) 1	-6.34295
poly(Subscription Length, 3) 2	14.93371
poly(Subscription Length, 3) 3	-26.90285
Charge Amount	-0.66984
Status: non-active	1.20880
Frequency of Use	-1.39272
Frequency of SMS	-0.80167

Example interpretation:

Call Failure:

For each additional call failure, the log odds of churning increase by 1.12131. This suggests that customers experiencing more call failures are significantly more likely to churn, reflecting a strong positive relationship between call failures and customer churn.

Charge Amount:

This coefficient indicates that as the charge amount increases (within its ordinal levels from 0 to 9), the log odds of churning decrease by 0.66984. Higher charges, possibly indicating higher usage or more premium services, appear to reduce the likelihood of churn.

6 Further Discussions

Misclassification cost is a big problem, due to the different losses if we do not apply retention strategy to churning groups or if we apply retention churning strategy to the non-churn group.

People may want to try neural network models or bayesian modelling techniques to the dataset to improve the accuracy, but due to time constraints, these methods are not included in this report.

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, 32, 14929-14962.

Y. Yang (2024). Group Lasso Penalized Learning Using a Unified BMD Algorithm.

M. Huang, T Brunsdon (2023). Applied Statistics Modelling Modules: Iranian Churn Data Analysis, University of Warwick.