Churn Analysis: An Application of Logistic Regression in a telecommunication company

Piermattia Schoch

March 13, 2019

Abstract

This project aims to exploit the problem of churn of customers from a telecommunications company, by means of an inferential analysis. The proposed methodology covers several phases: understanding the business; data processing and exploratory data analysis; implementing a logistic regression model for classification; evaluation of the most important variables affecting the probability of churn. The obtained results should be of great value for management and marketing departments.

1 Introduction

One of the more common tasks in Business Analytics is to try and understand consumer behaviour. By understanding the hope is that a company can better change this behaviour. In many industries it is more expensive to find a new customer then to entice an existing one to stay. Customer churn occurs when customers or subscribers stop doing business with a company or service. One industry in which churn rates are particularly useful is the telecommunications industry, because most customers have multiple options from which to choose within a geographic location. The aim of a "churn analysis" is to accurately identify the cohort who is likely to leave early enough so that the relationship can be saved.
In this report, we aimed to investigate the main reasons for churn. For this purpose, emphasising the fact that predictions are not the main objective, after processing the data, we implemented a logistic regression model, which gives the possibility to interpret the results much better than fancy algorithms popular nowadays.
The most important factors which are crucial for the customers to churn, are total minutes of calls made in the day per months,the numbers of calls to the customer services, the activation of the international plans with the corresponding amount of calls made and wheter the user is subscribed to the voice mail service.
The rest of this paper is organized as follows:
The next section defines the methodology for churn prediction,regarding different phases of the process. It presents a graphical inspection of the variables presented in the dataset and also describes the algorithms used for churn prediction. Section III presents the results of the variable selection procedure,

supported by many inferential tests, and finally test the assumption of validity of the logistic regression model applied. This report is concluded in Section IV.

2 Methodology for churn prediction

In order to find a possible solution to the problem of churn prediction i.e. successfully apply a statistical model to the available data, one needs a deep understanding of the project objectives and requirements from the telecommunications business perspective. The aim of the churn prediction is to identify the properties that make a customer churn in order to prevent it and retain the customer. To enable this, we consider customers that churned and analyze their data over a period while they still used the services of the telecommunications company.

2.1 Dataset

The initial dataset contains information about 3333 customers, with 22 (mainly numeric) attributes, that can be grouped in the following three categories:

* **Usage** : contains the information about the total numbers of calls expressed in minutes made per day in a month and their corresponding charge, calls made to customer service per month.

* **Contract attributes** : contain the attributes associated with the customer contract for a particular service such as the subscription of voice mail service, international calling.

* **Demographics** : contain the primary features of the customer such as the information about the State in which they lives, and its Area code.

2.2 Data Cleaning

The purpose of data cleaning is to reduce the number of inconsistent values, remove noise and incomplete entries and attributes. The dataset provided by the company has been already cleaned. Checking the descriptive statistics provided by the *summary()* function, nothing unusual has been found.

2.3 Data Exploration

Exploratory data analysis is an essential tool for receive a first hint about the data. Doing a great Exploratory analysis its a crucial step before getting in the construction of the model, since it permits to understand visually most of the insight that the data can provide.
Here there will be presented, firstly the proportion of customers that churned, then the relationship between the predictors and the response, respectively of each of three categories.

Figure 1 shows the percentage of customers who did not churn in the period of the analysis is $85,51\%$, while 14,49% of them had left the company.

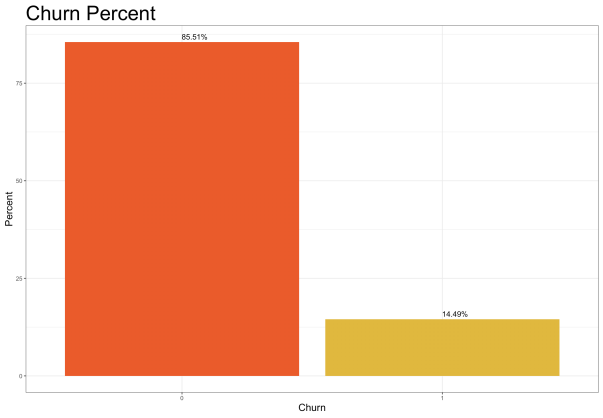

Figure 1: Customer churn proportion

A useful tip before performing the exploratory data analysis, is to check the distribution of each numerical predictor and their values of Pearson Correlation. On the right hand of the Figure 2, the Correlation values are highlighted only if they are relevant. A group of variable is exactly linearly correlated.

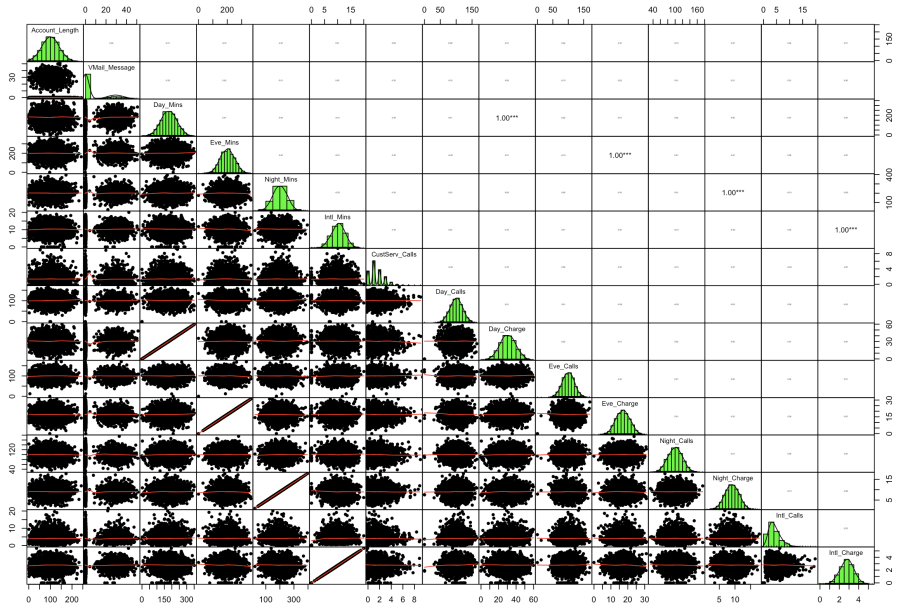

Figure 2: Bivariate scatterplot, Distribution, Pearson correlation

Looking the Figure 3 is evident that the two variables are redundant. This means that storing this amount of information in this dataset is not useful, since are exactly related. This behaviour permits to delete arbitrarly one of the two group of predictors. Since the behaviour of customer is the main focus of the analysis sounds better to delete all charge variables.

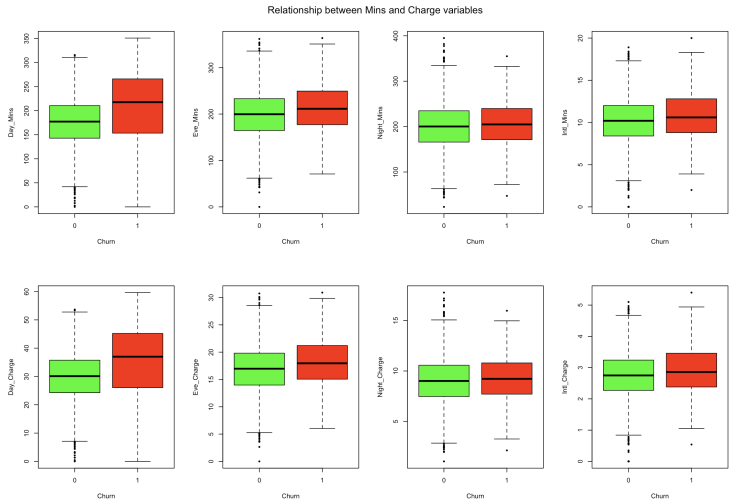

Figure 3: Relationship between Mins and Charge variables

Among all other variables concerning the usage categories (*Account_lenght* and *Day_calls,Evening_calls,Nights_calls* , boxplots shows that they have no influence on churning, except for *CustomerService_calls*, which shows a very interesting relation with the dependent variable: this trend is logical, given the fact that calling the customer service imply that complaints come up. This can be seen in Figure 4.

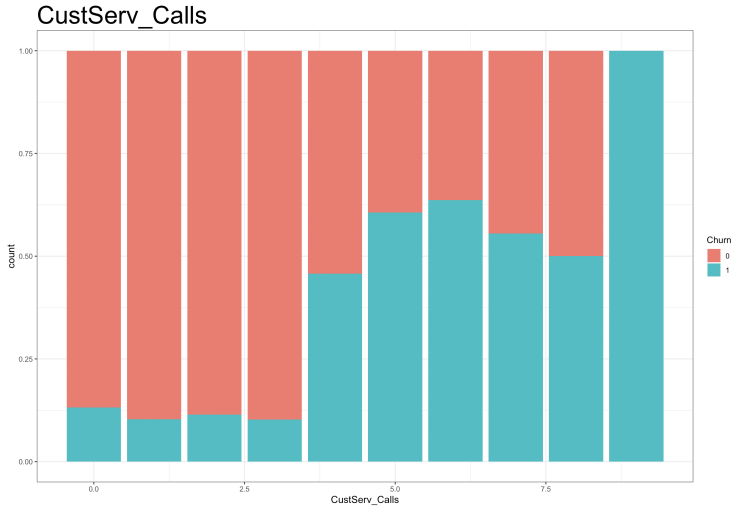

Figure 4: CustomerService calls and Churn proportion

Now it is helpful to understand how the additional services that the company offers, impact on the probability of churning. Figure 5 illustrate how customers

who decided to activate an international plan are more likely to churn, while customers that decided not to activate voice mail message tend to churn in lower proportion with respect to who did not use this service.

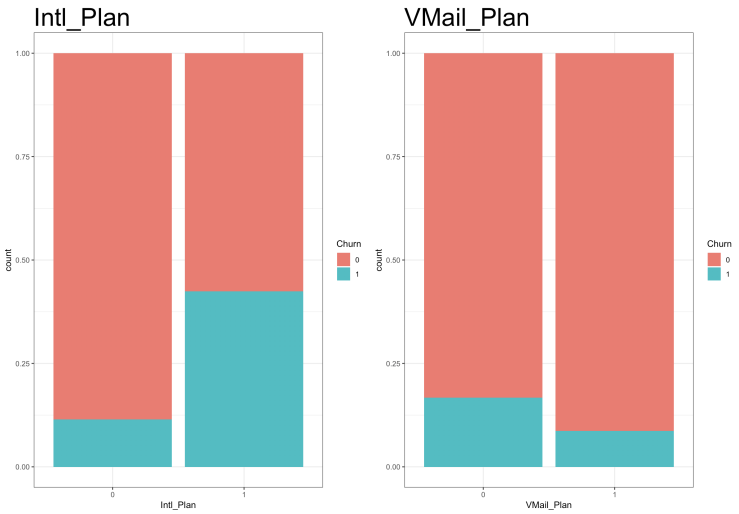

Figure 5: Additional services and Churn proportion

Figure 6 answer the question of how demographics variables impact on the churn ration. There are different state which have an higher proportion of customers that decided to leave the company, suggesting that this variable might be a useful predictor in our model.
On the other hand, Area Code, due to the fact that compress the information in State, c show more or less the same amount of people that churned in each of the three categories, with an higher rate in the last category (Area_510).
Being a male or female seems not to be very important: the proportion in very similar.

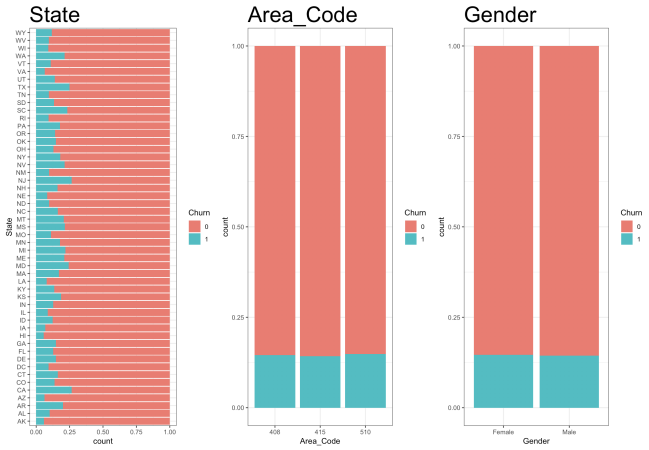

Figure 6: Demographics variables and Churn proportion

2.4 Data Transformation

Data transformation techniques can significantly improve the overall interpretability of the churn prediction, which i have seen while experimenting with potential transformations. The variable that has been enconde in different way are *VMail_Plan*, *VMail_Message* and *State*.
VMail_Plan is a binary categorical variable, which record the informations of customers who activated the voice mail service, and *VMail_Message* contains informations of the number of messages in user voicemail inbox. Due to the fact that the second variable is nested in the first, a extremely large of 0 are registred as can be shown in Figure 7.
The approach used in this situation is or including iteraction terms between the two variable, or to compress the information in a single variable for the sake of interpretability. The latter method is then preferred.
Therefore, *VMail_Plan* will be discard, keeping only *VMail_Plan*, compriming the values in an arbitrary way:

1. Low User: customers that have less than 25 voice messages
2. High User: customers that have more than 25 voice messages

As a result, Figure 8 represent the proportion of churn in each category. Setting a reference the category "No activation", it is possible to compare how the usage of this particular service impact on the probability of churning. Low and high users have a lower percentange of churned with respect to who did not use the service.

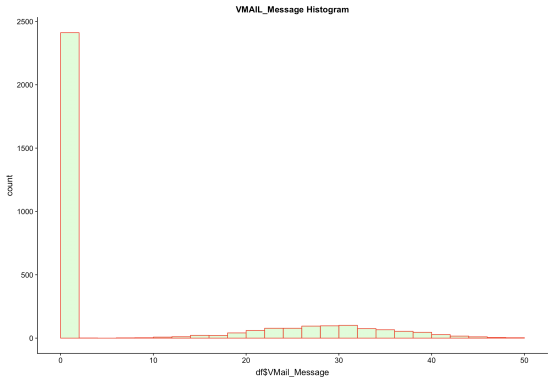

Figure 7: Histogram of VMail messages. Zero-inflated covariate

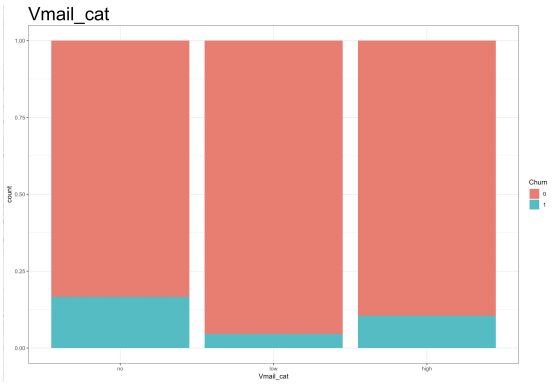

Figure 8: VoiceMail messages: No Active, Low user, High user

Another transformation has been made regarding the variable *State*, with the main aim to try to have an interpretable result.
Indeed this is a nominal variable, with 50 category, for which setting a reference level does not make sense, without any further business information. As a consequence, it has been transformed, according to the "US American Census", into 4 macro regions : Northeast, South, West, Middlewest. Generally it's best to choose as a reference category the one that makes interpretation of results easier. In this case is simply setted as South, the state which has most customer (1109). In Figure 9 it can be seen the proportion of churners in each macro-area.

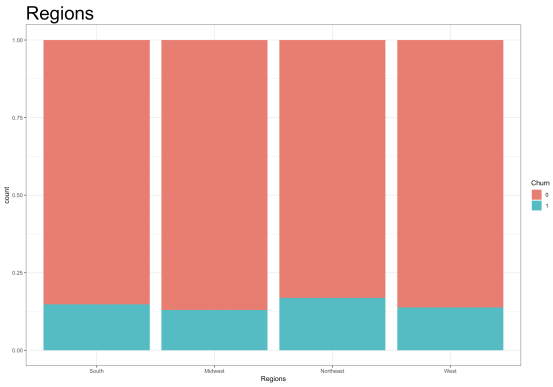

Figure 9: States in macro-areas: South, Midwest, West, North

3 Model Building

There are many techniques that have been proposed for customer churn prediction: decision tree, knearest neighbors algorithm, naive Bayes classifier and logistics regression. Among all of them, the one that permits the greatest interpretability of the results is certainly the logistic regression.

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 or 0, rappresenting the presence or absence of a certain event.
The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent variables.
Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest

$$\ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{1}$$

where p is the probability of presence of the characteristic of interest. In the next subsection will be illustrated the methodology to select the most important features that will be given as inputs to the logistic regression algorithm.

3.1 Features selection

Features selection refers to the process of selecting a subset of relevant attributes of a set of attributes. This reduces the number of input attributes to the learning algorithm, thereby significantly reducing time and resources required to train the algorithm, and most importantly permits to select a parsimonious model.
Here there will be applied a backward stepwise selection throught the *step* function in *R* software.
It begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.
Stepwise methods result in the creation of a set of models, each of which contains a subset of the p predictors. In order to implement these methods, we need a way to determine which of these models is best.
One possible solution is to use Aikake information criterion(AIC) or Bayesian information criterion(BIC), which are methods that permits to select among a set of models with different number of variables.
Knowing the fact that BIC statistics generally place a heavier penalty on models with many predictors, and hence results in the selection of smaller models than AIC, both criterion will be analyzed in Table 1.
The output in Table 1, show that both criterion choose the same variables.

Table 1: AIC vs BIC selection

	Dependent variable:	
	Churn	
	(1)	(2)
Eve_Mins	0.007^{***}	0.007^{***}
	(0.001)	(0.001)
CustServ_Calls	0.512^{***}	0.512^{***}
	(0.039)	(0.039)
Intl_Plan1	2.041^{***}	2.041^{***}
	(0.145)	(0.145)
Day_Charge	0.077^{***}	0.077^{***}
	(0.006)	(0.006)
Night_Charge	0.084^{***}	0.084^{***}
	(0.025)	(0.025)
Intl_Calls	-0.093^{***}	-0.093^{***}
	(0.025)	(0.025)
Intl_Charge	0.324^{***}	0.324^{***}
	(0.076)	(0.076)
VMail_catlow	-1.677^{***}	-1.677^{***}
	(0.322)	(0.322)
VMail_cathigh	-0.724^{***}	-0.724^{***}
	(0.157)	(0.157)
Constant	-8.069^{***}	-8.069^{***}
	(0.517)	(0.517)
Observations	3,333	3,333
Log Likelihood	-1,078.401	-1,078.401
Akaike Inf. Crit.	2,176.801	2,176.801
Note:	*p<0.1; **p<0.05; ***p<0.01	

3.2 Results

In this section will be analyzed the results obtained with the model selected in previous subchapter. In order to have a better idea of how the uncertainity and magnitude of the effect differs for these variables, it could be useful to scale the variable and plot their coefficients estimates with their confidence intervals (90%,95%) illustrated in Figure 10

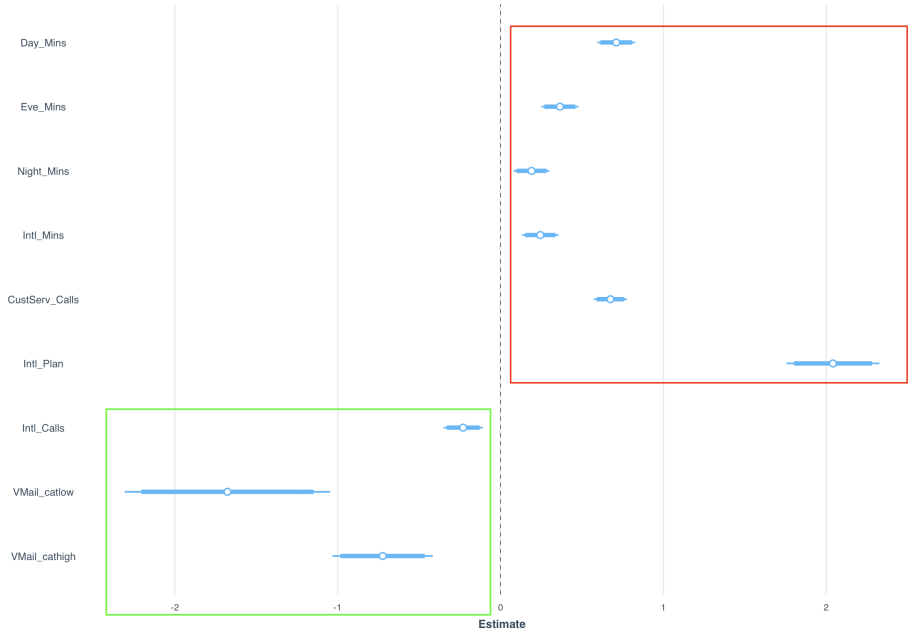

Figure 10: Scaled coefficients with CI at 90% and 95%

For exponential family models, it is interesting getting the exponentiated coefficients rather than the linear estimates. This are represented in the first column of Table 3, and are also called Odds Ratio. An odds ratio (OR) is a measure of association between an exposure and an outcome.
The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.
When a logistic regression is calculated, the regression coefficient (b1) is the estimated increase in the log odds of the outcome per unit increase in the value of the exposure. In other words, the exponential function of the regression coefficient ($e\hat{b}1$) is the odds ratio associated with a one-unit increase in the exposure.
The results are in Table 2 in the next page.

	coef	OR	2.5 %	97.5 %	pvalue
(Intercept)	-8.068	0.000	0.000	0.001	0.000
Day_Mins	0.013	1.013	1.011	1.015	0.000
Eve_Mins	0.007	1.007	1.005	1.009	0.000
Night_Mins	0.004	1.004	1.002	1.006	0.001
Intl_Mins	0.087	1.091	1.049	1.136	0.000
CustServ_Calls	0.512	1.669	1.547	1.804	0.000
Intl_Plan1	2.040	7.694	5.789	10.243	0.000
Intl_Calls	-0.093	0.911	0.867	0.956	0.000
VMail_catlow	-1.677	0.187	0.095	0.337	0.000
VMail_cathigh	-0.724	0.485	0.354	0.655	0.000

In the following list will be evaluated the results, in term of percentange change in the Odds. For each percentange change all the other variable must be hold at a fixed value.

1. Day_Mins: 1.3% increase in the odds of churning for one unit increase in minutes used during the day
2. Eve_Mins: 0.7% increase in the odds of churning for one unit increase in minutes used during the evening
3. Night_Mins: 0.4% increase in the odds of churning for one unit increase in minutes used during the night
4. Intl_Mins: 9.1% increase in the odds of churning for one unit increase in minutes used internationally
5. CustServCalls: 66,9% increase in the odds of churning for one unit increase in customerservice calls made
6. IntlPlan1: the odds of who has activated the international plan are 669,4% higher than who did not activated the service.
7. Intl_Calls: 8.9% decrease in the odds of churning for one unit increase in international calls made
8. VMail_catlow: the odds of who has activated the voice mail service and have received less than 25 messages are 81,3% lower than who did not activated the service
9. VMail_cathigh:the odds of who has activated the voice mail service and have received more than 25 messages are 51,5% lower than who did not activated the service

3.3 Godness of Fit

After fitting a model to the observed data, one of the next essential steps is to investigate how well the proposed model fits the observed data.

Unlike linear regression with ordinary least squares estimation, there is no R2 statistic which explains the proportion of variance in the dependent variable that is explained by the predictors. However, there are a number of pseudo R2 metrics that could be of value. Most notable is McFaddens $R\hat{2}$,which ranges from 0 to just under 1, with values closer to zero indicating that the model has no predictive power.

	bic_model	full_model
llh	-1078.416	-1074.184
llhNull	-1379.147	-1379.147
G2	601.460	609.925
McFadden	0.218	0.221
r2ML	0.165	0.167
r2CU	0.293	0.297

In Table 3 are computed:

1. llh: Log-likelihood from the fitted model
2. llhNull: The log-likelihood from the intercept-only restricted model
3. G2: Minus two times the difference in the log-likelihoods
4. McFadden: McFadden's pseudo r-squared
5. r2ML: Maximum likelihood pseudo r-squared (Cox and Sheel)
6. r2CU: Cragg and Uhler's pseudo r-squared (Nagelkerke's pseudo r-squared)

The results show an McFadden value of 0.21 which can be classificate as moderate, even in there is no standard rule. Differently to the values taken from R2 in multiple linear regression tends to be lower, and moreover comparing the result with the one of the Full_Model (that include all the variables) are almost identical, indicating that with the variable selection made we are not losing informations and predictive power.

It has also been applied the "Hosmer-Lemeshow test" (X-squared = 18.918, p-value: 0.01531) that indicate that there is not enough evidence to say it's a good fit.

3.4 Model Diagnostic

Generalized linear models are built on some probabilistic assumptions that are required for performing inference on the model parameters.

3.4.1 Linearity

Linearity between the logit of Y and the predictors $X_1, ..., X_p$, is the building block of generalized linear models. If this assumption fails, then all the conclusions we might extract from the analysis are suspected to be flawed. Therefore it is a key assumption.

In order to check it, residual vs fitted values plot should be used. Under the linearity we expect that there is no trend in the residuals. At best, the trend is a horizontal straight line without curvature. The default residual for generalized linear model is Pearson residual. Figure 11 plots Pearson's residual against predictors one by one and the last plot is against the predicted values (linear predictor). The linearity assumption is respected.

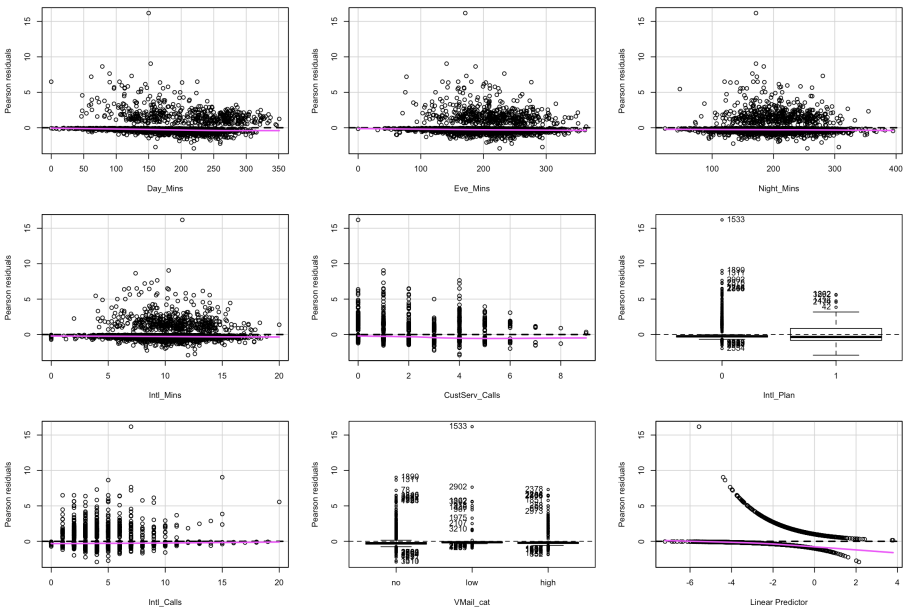

Figure 11: Pearson residuals vs. fitted values plots (first row) for datasets (second row) respecting the linearity assumption in logistic regression.

3.4.2 Multicollinearity

Multicollinearity can also be present in generalized linear models. Due to this, if two or more predictors are highly correlated between them, the fit of the model will be compromised since the individual linear effect of each predictor is hard to distinguish from the rest of correlated predictors. Then, a useful way of detecting multicollinearity is to inspect the VIF of each coefficient. The result are shown in Table 4, that shows that there are no correlated predictors, since every value is close to 1, indicating absence of multicollinearity.

	GVIF	Df	GVIF^(1/(2*Df))
Day_Mins	1.0475	1.0000	1.0235
Eve_Mins	1.0286	1.0000	1.0142
Night_Mins	1.0157	1.0000	1.0078
Intl_Mins	1.0137	1.0000	1.0068
CustServ_Calls	1.0866	1.0000	1.0424
Intl_Plan	1.0678	1.0000	1.0334
Intl_Calls	1.0101	1.0000	1.0051
VMail_cat	1.0204	2.0000	1.0051

4 Conclusion

The telecommunication industry in the recent years is a subject of major changes and from a fast-growing industry has come to a state of saturation accompanied with strong competitive market. Customers starve for better services and prices, while their requirements are extremely complex and difficult to understand. In order to cope with this problem, through this in analysis has been possible to identify, the most important factor that influence the probability of churn for those customer. This analysis has revealed that customer that are more active, tend to change their telecommunication company more easily. Especially those who activated the international plan, have an extremely high chance of churning respect to ones who choose not to use this service. Moreover, the effect of calling during the day is higher than those registred during the evening and the night, behaviour that bring to mind that difference in demographics, habits play a role. On the other hand, the possibility of receiving a voice mail message turns out to be a positive factor for the company. Who decided to use this characteristic, have a lower probability to churn, but also here is possible to see that customer that are more active in using the service provided by the company, might change provider more frequently. As further analysis, it could be interesting to include more demographics variables, such as age, income, occupation, informations like how frequently they come to store, navigate the website in their personal area, the reason of complain and informations about how frequently they purchase offers and in which way they pay.

5 Bibliography

1. Rui Miguel Forte (2015). Mastering predictive analytics with R
2. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition). Springer
3. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.