

MIS-64036: Business Analytics

Group 4 Project Report

Group Members

1. Spence, Steven (sspenc12)
2. Marjanovic, Marianne (mmarjan1)
3. Omeike, Stanley (someike)
4. Qutub, Manar (mqutub)

1.0 Project Goal

ABC Wireless Inc. has invited a study into its customer's churn issue with the aim of determining the key factors from existing data mine that influence customer loss.

2.0 Project Method

This project will use historical data from ACB Wireless Inc. to build a model that can predict/identify their customers who are likely to churn.

3.0 Literature Review: Customer Churn: Theoretical Context

Customer churn in the telecoms sector has been defined as customer erosion (Yang, and Chiu, 2006). This phenomenon often involved switching from one service provider to another (Hadden,Tiwari , Roy and Ruta ,2005). Churn in the highly competitive telecoms sector causes revenue loss and degraded financial performance. As a result, the goal of any organization is to minimize churn and improve customer loyalty.

Three types of customer churn has been proposed in literature (Yang and Chiu ,2006): Involuntary churn(This occurs when customers fail to pay their bill and as a result, the provider terminates service), Inevitable churn (This occurs when customers die or migrate resulting in omitting customer from market completely) and Voluntary churn (This occurs when customers prefer to switch to another operator because of more value).

A quick look at literature identifies several well researched factors that bear strong correlation to churn. The following are seven of such factors;

A. Service price: Service price is defined as the amount of money that a customer pays for services and the goal of consumers is to minimize this while retaining or even maximizing

service benefits. As a result, customers tend to be attracted to the competitors with lower prices. service providers that effectively drive down service cost to the customer increase customer attraction and reduce customer churn. Anuwichanont, (2011) therefore argues that higher prices have negative effects on customer purchases and positive effects on customer churn. This corroborates several other research findings that perceived fair price is an influential factor on churn. (Jiang and Rosenbloom, 2005).

B. Switching cost: White and Yanamandram, (2007) define Switching cost as the economic cost incurred by the customer as a result of switching from one service provider to another. Shi, Zhou, And Liu, (2010) extend this cost structure to include physical, emotional and time dimensions. Therefore, switching service providers can trigger immense tangible and intangible costs to the consumer and can act to discourage switching. Hejazi Nia, (2013) conclude that unsatisfied customer may elect to remain with a service provider if the cost of switching is perceived to be high. Therefore, we infer that switching cost has negative effects on customer churn.

C. Competitors with superior technology: Service providers with superior technology are competitors that leverage advanced technologies to improve service quality and features (Hejazi Nia ,2013). Competitors with state-of-the-art technologies and fewer prices magnetize customers. Service providers with improved service quality and features encourage customers to switch easily from other service providers lacking these benefits (Jones and Sasser 1995). Therefore, as satisfaction decreases, customers become eager to change their service provider (Jones, Mothersbaugh and Beatty ,2000).

D. Quality: Quality is the difference between customer perception and experience of service (Grönroos 1997). Parasuraman (1998) extends this definition to include the difference between actual services and the promised services. Therefore, in service organizations, quality is dependent on the extent to which the service promise meets customer expectations. Ahn, et.al, (2006) further refines this for the telecoms companies by stating that service quality refers to call quality consisting of audio, video, and text services provided by operator during a call. Therefore, quality can be seen to affect customer churn.

E. Satisfaction: Oliver (1980) defines customer satisfaction as the perceived value minus expectations of the customer. if received value is equal to expectations, satisfaction is created. Jamal and Naser (2002) further refines this view by stating that customer

satisfaction is the overall customer attitude toward a product or service after using it. Increased satisfaction prevents customer churn.

F. Security concern: Gappert, (2002) highlights security concern as a signification factor in customer churn. hejazi nia, (2013) then clarifies that security concerns refers to the fear of losing data or personal information as a result of using the services of a service provider. cahill, (2007) provides a framing of trust and confidence as an indicator of this. The less the trust and confidence, the higher the potential for customer churn

G. Advertising: advertising propagates brand philosophy and appeal (Kotler and Armstrong, 2000). Hejazinia, (2013) argue that increased advertising attracts new customers and improves customer loyalty. Therefore, increased advertising increases loyal customers and prevent customer churn.

In summary, we find that customer churn as a phenomenon has been well researched and several factors that influence this phenomenon have also been identified in literature. Several approaches have been used to study the effect of these factors on churn. In all cases, corporate executives focus on identifying factors that show strong correlation with churn and then seek to implement programs that either minimize or optimize the effects of these factors to drive down churn.

In the next section, we explore ABC Wireless INC's data set and set out our approach to data exploration and model building to surface insights that will enable ABC Wireless to improve its effectiveness at minimizing churn.

4.0 ABC Wireless INC Data Set Structure

ABC Wireless INC's dataset summary view reveals a dataset with 19 predictors of churn in its customer base;

1. state (categorical),
2. account_length,
3. area_code,
4. international_plan (yes/no),
5. voice_mail_plan (yes/no),
6. number_vmail_messages,
7. total_day_minutes,
8. total_day_calls,
9. total_day_charge,

10. total_eve_minutes,
11. total_eve_calls,
12. total_eve_charge,
13. total_night_minutes,
14. total_night_calls,
15. total_night_charge,
16. total_intl_minutes,
17. total_intl_calls,
18. total_intl_charge
19. number_customer_service_calls.

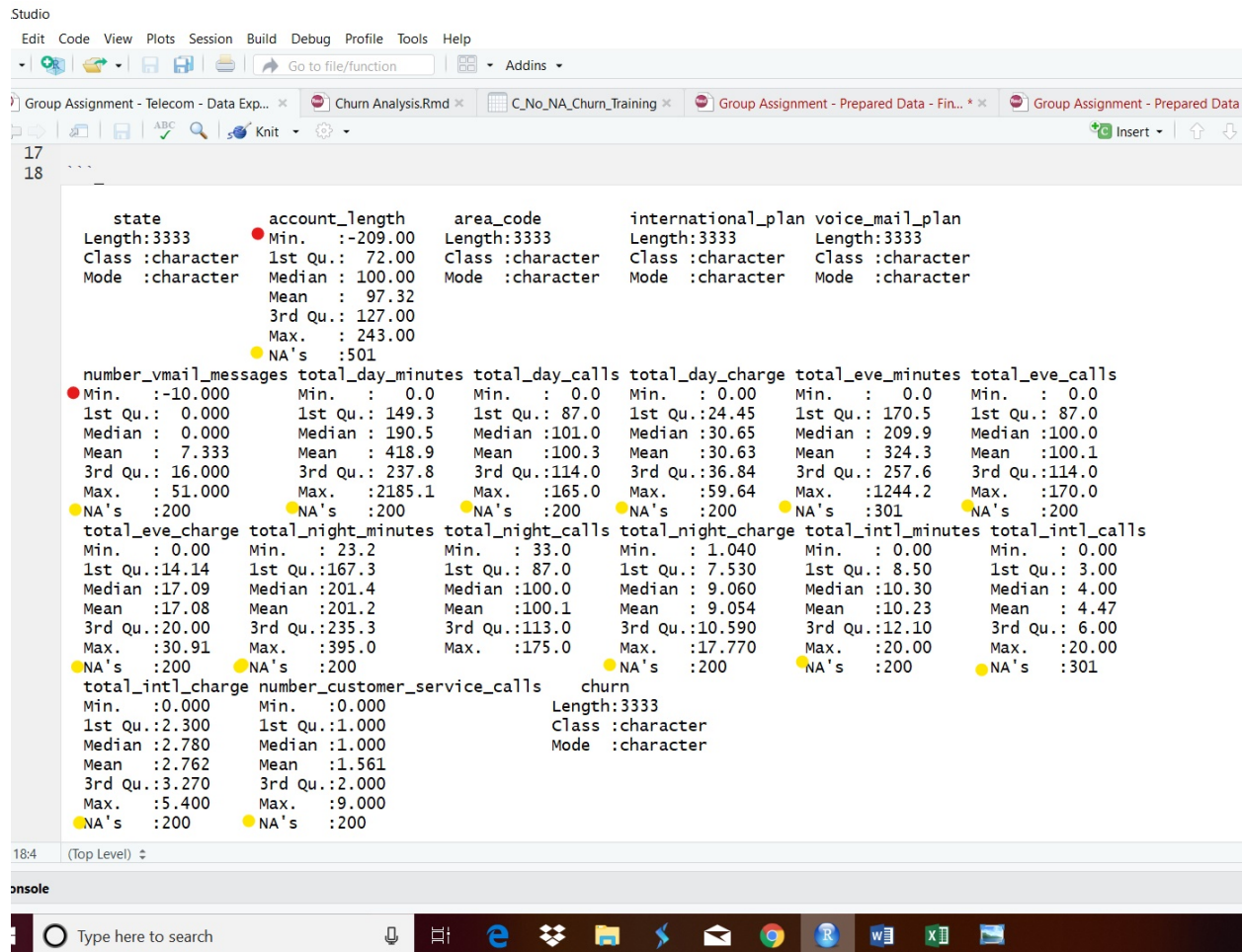
In section two we conduct multi-factor correlation analysis to establish the correlation of factors/combination of factors with Churn and the extent to which each of the tested models predict churn using the R^2 values of the model.

5.0 Pre-Model Development Factor Influence Analysis

1. Poor Customer services will result in Customer churn.
2. Call Charge will influence Customer churn. Literature refers to this as Cost of Service.
3. International Plan Customers are most-likely to churn if we assume that majority of them are Non-Resident citizens.
4. Frequency of customer service calls is a proxy for weak quality perception and will positively influence churn. Literature refers to this as Service Quality.
5. Customer churn will be stronger among lower income, low volume groups. These groups will use less service features and avoid international calls.
6. The longer the account length the longer customer tenure and the stronger the negative influence on churn.

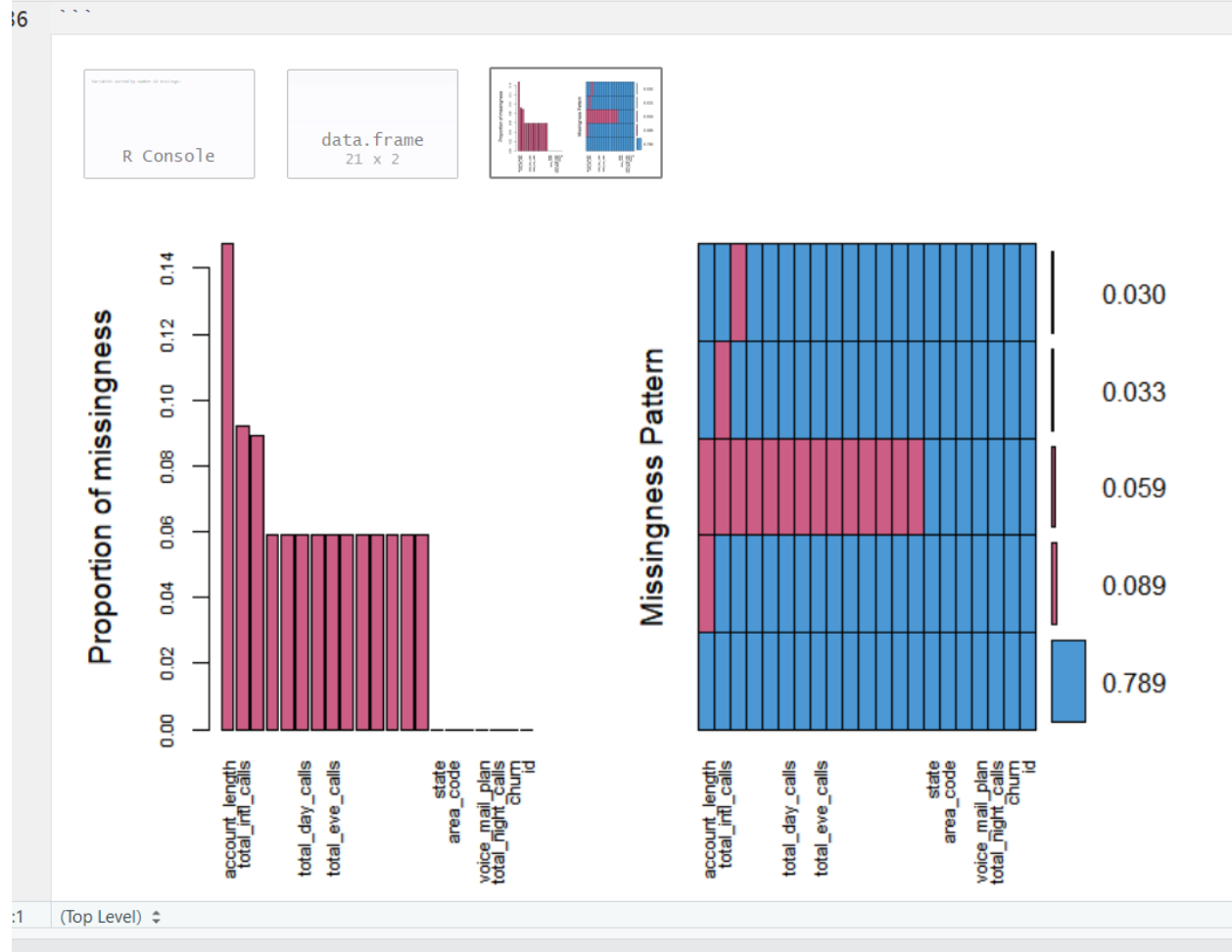
6.0 Data Summary Statistics and Initial Analysis:

After identifying the variables provided to us in the data set, we wanted to get a feel for the data's overall quality and cleanliness. This was done by calling the data set summary to provide a cursory overview for each variable as shown below.



```
17
18
state      account_length area_code international_plan voice_mail_plan
Length:3333 Min. : -209.00 Length:3333 Length:3333 Length:3333
Class :character 1st Qu.: 72.00 Class :character Class :character Class :character
Mode :character  Median : 100.00 Mode :character Mode :character Mode :character
                Mean  : 97.32
                3rd Qu.: 127.00
                Max.   : 243.00
                NA's   :501
number_vmail_messages total_day_minutes total_day_calls total_day_charge total_eve_minutes total_eve_calls
Min. : -10.000 Min. : 0.0 Min. : 0.0 Min. : 0.00 Min. : 0.0 Min. : 0.0
1st Qu.: 0.000 1st Qu.: 149.3 1st Qu.: 87.0 1st Qu.:24.45 1st Qu.: 170.5 1st Qu.: 87.0
Median : 0.000 Median : 190.5 Median :101.0 Median :30.65 Median : 209.9 Median :100.0
Mean : 7.333 Mean : 418.9 Mean :100.3 Mean :30.63 Mean : 324.3 Mean :100.1
3rd Qu.: 16.000 3rd Qu.: 237.8 3rd Qu.:114.0 3rd Qu.:36.84 3rd Qu.: 257.6 3rd Qu.:114.0
Max. : 51.000 Max. :2185.1 Max. :165.0 Max. :59.64 Max. :1244.2 Max. :170.0
NA's :200 NA's :200 NA's :200 NA's :200 NA's :301 NA's :200
total_eve_charge total_night_minutes total_night_calls total_night_charge total_intl_minutes total_intl_calls
Min. : 0.00 Min. : 23.2 Min. : 33.0 Min. : 1.040 Min. : 0.00 Min. : 0.00
1st Qu.:14.14 1st Qu.:167.3 1st Qu.: 87.0 1st Qu.: 7.530 1st Qu.: 8.50 1st Qu.: 3.00
Median :17.09 Median :201.4 Median :100.0 Median : 9.060 Median :10.30 Median : 4.00
Mean :17.08 Mean :201.2 Mean :100.1 Mean : 9.054 Mean :10.23 Mean : 4.47
3rd Qu.:20.00 3rd Qu.:235.3 3rd Qu.:113.0 3rd Qu.:10.590 3rd Qu.:12.10 3rd Qu.: 6.00
Max. :30.91 Max. :395.0 Max. :175.0 Max. :17.770 Max. :20.00 Max. :20.00
NA's :200 NA's :200 NA's :200 NA's :200 NA's :200 NA's :301
total_intl_charge number_customer_service_calls churn
Min. :0.000 Min. :0.000 Length:3333
1st Qu.:2.300 1st Qu.:1.000 Class :character
Median :2.780 Median :1.000 Mode :character
Mean :2.762 Mean :1.561
3rd Qu.:3.270 3rd Qu.:2.000
Max. :5.400 Max. :9.000
NA's :200 NA's :200
```

From this summary, we were able to evaluate that 2 variables, depicted with red dots, account_length and number_vmail_messages contained inappropriate negative values and 14 variables, depicted with the yellow dots above, contained at least 200 rows of missing data (NA's). Looking further into the missing data, using the VIM package in R, we ran md.patterns which generated the graphic below which shows the layout pattern of the NA data. This graphic tells us that all 14 missing variables share 200 common rows and 3 of those 14 variables have an additional 101 rows each of missing data.



All the information ascertained so far in this information seeking process was very useful to us in making our decision to treat the rows of data containing the variables with negative values and the rows containing missing data. But, before treating the missing data and negative values, we went ahead and constructed histograms for each numeric variable. We wanted to see the distributions for these variables before and after data treatment. Our suspicion was that the distributions prior to treatment would be a bit skewed and distorted showing the effects of negative data and missing data. The histograms after data treatment confirmed this suspicion as the distributions for the variables with problematic data were normalized.

7.0 Factor Correlation Analysis and Interpretation:

Treatment of the Data

In the data treatment phase there are 2 types of data that needed to be treated in our dataset. First, negative value data in the `account_length` and second, the missing data for the following 14 variables:

1. `account_length`,
2. `number_vmail_messages`,
3. `total_day_minutes`,
4. `total_day_calls`,
5. `total_day_charge`,
6. `total_eve_minutes`,
7. `total_eve_calls`,
8. `total_eve_charge`,
9. `total_night_minutes`,
10. `total_night_charge`,
11. `total_intl_minutes`,
12. `total_intl_calls`,
13. `total_intl_charge`
14. `number_customer_service_calls`.

Before addressing the negative data value situation, we assumed that these values were incorrectly entered as negative and that the absolute value for these variables are the correct value. So, to correct this data entry error, we used the `abs()` function in R to convert the negative `account_length` values to their absolute values.

Next, we used the MICE package in R to impute values for the missing values in the data set. MICE does this by determining the best method suitable for continuous incomplete variables. The method that MICE used in this case is “pmm” or predictive mean matching. PMM finds a set of observed values with the closest predicted mean as the missing one and imputes the missing values by a random draw from that data set. MICE uses all available variables in its’ imputation model including categorical values.

8.0 Model Building and Testing:

Next, we built the Logistic Regression Model to predict which measurable variables are important and most significant to cause the churn.

Call:

```
glm(formula = churn ~ international_plan + voice_mail_plan +
    total_day_charge + number_customer_service_calls + percent_day_calls +
    percent_eve_calls + percent_night_calls + percent_day_minutes +
    percent_eve_minutes + percent_night_minutes + percent_day_charge,
    family = "binomial", data = Imputed.Churn.Training)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0953	-0.4844	-0.3067	-0.1720	3.2572

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	24.0700	15.6295	1.540	0.123551
international_planyes	2.1917	0.1758	12.469	< 2e-16 ***
voice_mail_planyes	-0.9654	0.1744	-5.534	3.13e-08 ***
total_day_charge	0.2254	0.0170	13.256	< 2e-16 ***
number_customer_service_calls	0.5579	0.0467	11.947	< 2e-16 ***
percent_day_calls	34.3379	9.1800	3.741	0.000184 ***
percent_eve_calls	33.9891	9.2262	3.684	0.000230 ***
percent_night_calls	34.0613	9.2457	3.684	0.000230 ***
percent_day_minutes	-58.2849	12.8200	-4.546	5.46e-06 ***
percent_eve_minutes	-61.2092	13.4031	-4.567	4.95e-06 ***
percent_night_minutes	-59.5213	13.3765	-4.450	8.60e-06 ***
percent_day_charge	-18.0985	1.9308	-9.373	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2048.3 on 2505 degrees of freedom

Residual deviance: 1515.8 on 2494 degrees of freedom

(160 observations deleted due to missingness)

AIC: 1539.8

Number of Fisher Scoring iterations: 6

After taking a closer look at the Z-Value, the most significant values are:

- International Plan
- Voicemail Plan
- Total Day Charges
- Number of Customer Services Calls
- Percent Day Calls
- Percent Eve Calls
- Percent Night Calls
- Percent Day Minutes
- Percent Eve Minutes
- Percent Night Minutes
- Percent Day Charge

Finally, we implemented a Performance Matrix to back up the model and to explain more the variable importance.

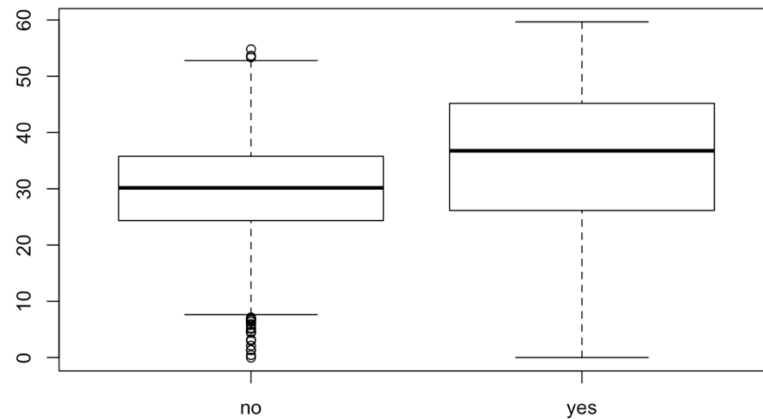
- Here we have the International Plan Matrix, it shows that customers with an international plan appear to churn 4x the rate of customers without an international plan.

international_plan <fctr>	count <int>	percent_churn <dbl>
yes	323	42.41486
no	3010	11.49502

- For the next variable Voice Mail Plan, we see that customers without a voicemail plan have a churn rate 2x more than customers with a voicemail plan.

voice_mail_plan <fctr>	count <int>	percent_churn <dbl>
no	2411	16.71506
yes	922	8.67679

- Next, this graph shows that customers that churn have a higher total day charge rate on average. The median total day charge is approximately 22.5% higher for customers that churn.

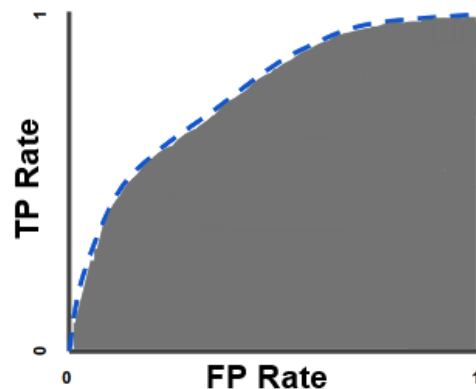


- After reviewing the Number of Customer Services Calls, the matrix shows that customers are much more likely to churn if they've called customer service 4 times or more. Churn percentage increases from 10% to at least 40% after the 3rd customer service call.

number_customer_service_calls <int>	count <int>	percent_churn <dbl>
0	695	13.52518
1	1172	10.32423
2	772	11.52850
3	425	10.82353
4	167	43.71257
5	68	57.35294
6	21	66.66667
7	10	50.00000
8	2	50.00000
9	1	100.00000

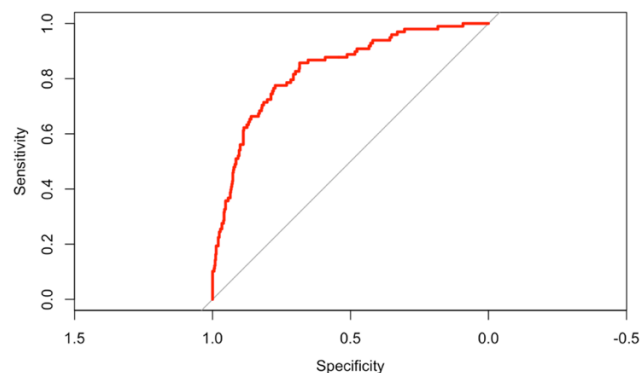
9.0 Model Fit Statistics

To assess the fitness of our model we will be using AUC (Area Under the Curve). AUC measures the area under the ROC curve which essentially plots true positive observations against false positive observations. The result is a measure of discrimination.



AUC results range from 0.0 to 1.0. A model that is **100% correct** scores a 1.0. Likewise, a model that scores a 0.0 is **100% incorrect**. In the real world, it is unlikely that that a model will have a perfect score of 1.0. Therefore, it is assumed that any score above a threshold of 85% is very strong and is taking into account all variables that are relevant and influential in the predicted outcome.

A screenshot of the logistic regression model performance is shown below from the “pROC” package in RStudio. This plots out the sensitivity performance values versus the specificity performance values to obtain an AUC value. This AUC value can then be compared with other model’s values to see which model captures the most variability, resulting in a more accurate model. The model used in this analysis was found to have an AUC value of approximately 85%.



10. Study Findings

At the very beginning we had some assumptions before even starting the Data Analysis. One of the assumptions is that we were very confident that International Plan Customers are most-likely to churn assuming most of them are Non-Resident citizens. These Customers are in the country for a specific goal and reason, after they fulfil this purpose or find a better opportunity in a different country ABC company is going to lose them. We looked at this variable as why these customers will stick with one telecom provider until their time is finished, we assumed that they'll keep churning on every company until they find the company that meets their requirements. Another important thing to highlight is that these customers don't have a SSN so they can easily churn without paying their bills leaving no hard trail for the company to reach out to them.

Another assumption was no matter how great the company is, overall success strongly depends upon customer service. If the customer service was poor then customers won't stay loyal to this company, they'll churn and make other customers churn with them. These assumptions that we've made require all the support of our findings to hold true, otherwise we can't say for sure that they are significant. We also included other important variables to cause churn at the initial analysis. Other important variables to cause churn were discovered only after the analysis, we couldn't eliminate any variable until we were done with the study.

After we finalized the analysis along with the Performance Matrix and everything else was done, we were very pleased that our initial assumptions were confirmed by the Data Analysis. The following is a Business Insight Report made for ABC company to help maintain customers:

Significant Variables	Business Insight
International Plan	Customers appear to churn 4x the rate of customers without an International plan.
Number of Customer Services Calls	Customers are much more likely to churn after calling customer service 4 times or more. Churn % increases from 10% to at least 40% after the 3rd customer service call.
Voicemail Plan	Customers without a Voicemail plan have a churn rate of 2x the customers with a Voicemail plan.

Total Day Charges	Customers that churn have a higher total day charge rate on average. The median total day charge is approximately 22.5% higher for customers that churn.
Percent Day Minutes	Customers churning have a slightly higher day usage percentage than customers that do not (~1%)
Percent Night Minutes	Customers churning have a slightly less night percentage of mins used than customers that do not churn (~1%)
Percent Day Charge	Customers churning have a slightly higher percent of day charges than those customers not churning
Percent Day Calls	Logistic regression model finds this variable to be significant, but multivariate plotting of this variable did not find any obvious patterns between churn.
Percent Evening Calls	Logistic regression model finds this variable to be significant, but multivariate plotting of this variable did not find any obvious patterns between churn.
Percent Night Calls	Logistic regression model finds this variable to be significant, but multivariate plotting of this variable did not find any obvious patterns between churn.
Percent Evening Minutes	Logistic regression model finds this variable to be significant, but multivariate plotting of this variable did not find any obvious patterns between churn.

11.0 Conclusion and Recommendation

Customer churn in the telecoms industry has been a key factor driving competitive advantage. Our study of ABC Wireless and its dataset supports most of the findings in literature; From our analysis, leading causes of churn are the quality of service reflected in the number of service calls (z Value = 11.9) , International Plan (z value= 12.5), Voicemail Plan (z value = -5.5), Total Day Charges (z value = 13.3) and Percent Day Charge (-9.4).

We argue that these factors and their significance can be explained in terms of service quality and service price. Our findings show that day service charges are a strong predictor of churn. This is supported by other findings in literature (Anuwichanont, 2011; Jiang and Rosenbloom, 2005) that suggest that higher prices have negative effects on customer purchases and positive effects on customer churn. Our findings show that the median total day charge for customers that churn is about 22.5% higher. This represents increased service cost which ABC wireless can target in order to reduce churn. This pattern also explains why there is 42% churn among customers with International call plans.

We also find that the number of customer calls to customer service is a strong predictor of churn. The higher the number of calls the higher the churn. Generally speaking, the more a customer has to call into customer service with issues the higher the level of dissatisfaction with the quality of service. ABC Wireless Telecoms dataset shows that customers are much more likely to churn if they've called customer service 4 times or more. Churn percentage increases from 10% to at least 40% after the 3rd customer service call. Therefore, quality of service is negatively affected by increasing number of calls to customer service. This is consistent with literature Ahn, et.al, (2006). the Number of Customer Services Calls, the matrix

In relation to our Pre-model development hypothesis we reach the following conclusions;

1. Poor Customer services will result in Customer churn.: *This is supported by our findings; Poor customer service will result in customer churn as shown by the z-value of customer service calls.*
2. Call Charge will influence Customer churn. Literature refers to this as Cost of Service.: *This is supported by our findings. The churn rate among customers with higher day time charges were significantly high.*
3. International Plan Customers are most-likely to churn if we assume that majority of them are Non-Resident citizens.: *This is supported by our findings since international calls are charged at a higher rate and the churn amongst this customer group is about 42%.*
4. Frequency of customer service calls is a proxy for weak quality perception and will positively influence churn. Literature refers to this as Service Quality.: *Our findings support this hypothesis and further shows that Churn increased significantly as the number of calls to customer service increased.*
5. Customer churn will be stronger among lower income, low volume groups. These groups will use fewer service features and avoid international calls. *We could not establish this finding based on the provided dataset.*
6. The longer the account length the longer customer tenure and the stronger the negative influence on churn. *We could not establish this finding based on the provided dataset.*

Recommendations

ABC Wireless needs to implement a customer segmentation program to target its international and day time customers with cost-saving service packages aimed at reducing the perceived cost of service by this market segment. This program needs to assure also that service quality is strengthened through further analysis of customer service call issues and the service resolution process.

The customer service call centers would benefit from a process transformation to ensure that customer complaints and calls are resolved within the first call and ensure that customers do not need to call three or more times to have their issues resolved. Unfortunately, the provided dataset does not present sub datasets of customer complaints to enable either a semantic analysis of customer perception of quality or an analysis of service factor perceptions among customers. We recommend that ABC Wireless either provide the records of its customers complaints or actively collect this data for future analysis.

References

1. Ahn, J., Hana, S., & Lee, Y. (2006). "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry". *Telecommunications Policy*, 30, 552–568.
2. Anuwichanont, J. (2011), "The Impact of Price Perception on Customer Loyalty in the Airline Context", *Journal of Business and Economics Research*, 9, 37-49.
3. Cahill, D.L. (2007). "Customer Loyalty in Third Party Logistics Relationships: Findings from studies in Germany and USA", New York: Physica Verlag Heidelberg Publication.
4. Gappert, C. (2002). "Customer Churn in the Communications Industry", A KPMG LLP white paper, U.S member of KPMG International.
5. Grönroos, C. (1997). "From Marketing Mix to Relationship Marketing - Towards A Paradigm Shift in Marketing", *Management Decision*, 35, 322-339.
6. Hadden, J., Tiwari, A., Roy, R. & Ruta, D. (2005). "Computer assisted customer churn management: State-of-the-art and future trends", *Computers & Operations Research*, 44 (10), 2902-2917.
7. Hejazi Nia, R. (2013). "Identify the variables affecting the Telecommunication company's customers ", Final thesis in master's degree, university of sistán and bluchestan.
8. Jamal, A., Naser, K. (2002). "'Customer Satisfaction and Retail Banking: An Assessment of son of the Key Antecedents of Customer Satisfaction in Retail Banking", *International Journal of Bank Marketing*, 20(4), 146-160
9. Jiang, P. & Rosenbloom, B. (2005). "Customer intention to return online: price perception attribution-level performance, and satisfaction unfolding over time", *European Journal of Marketing*, 39(1/2), 150-174.
10. Jones, T.O. & Sasser, W. E. (1995). "Why satisfied customers defect?", *Harvard Business Review*, 73(6), 88-99.
11. Jones, M. A., Mothersbaugh, D.L & Beatty, S.E. (2000). "Switching barriers and repurchase intentions in services", *Journal of Retailing*, 76(2), 259–274.
12. Kotler, P. & Armstrong, G. (2000). "the principle of marketing", prentice and hall, international edition .
13. Oliver, R. L. (1980). "A cognitive model of the antecedents and consequences of satisfaction decisions", *Journal of Marketing Research*, 91, 462-464.
14. parasuraman, A. (1998). "Customer Service in business –to- business markets: an agenda for research", *Journal of Business & Industrial Marketing*, 13(4/5), 309-321.
15. SHI, W. H., ZHOU, W. & LIU, J.Y. (2010). "Analysis of the influencing factors of users' switching intention in the context of one-way mobile number portability", *The Journal of China Universities of Posts and Telecommunications*, 17, 112-117.
16. White, L. & Yanamandram , V. (2007). "A model of customer retention of dissatisfied business services customers", *Managing Service Quality*, 17, 298-316.
17. Yang, L. S. & Chiu, C. (2006). "Knowledge Discovery on Customer Churn Prediction", *Proceeding of the 10th WSEAS international conference on applied mathematics*, Dallas, Texas, USA.