Telco ABC. Report

Marketing Analytics Assignment

GROUP V

Gena Keenan	40208341
James Taylor	40312691
Linghe Tian	40320019
Vasileios G. B.	40314803
Zhiying Zhu	40309491

Word count: 4835

Table of Contents

Executive Summary

Telco ABC is a telecommunications company that has commissioned this report to devise a new communication strategy to improve the levels of our customer retention. Telco has been facing an increase in customer churn in recent years and are in need of a new strategy to reverse or at least mititgate this trend.

To reduce customer churn, this report will analyse the consumer data given and draw up recommendations. As discussed later in this report, it is recommended that there should be a focus on our target customer of families and subsequently offer a 'family bundle'. Success in improving retention will drive from ensuring the needs of our customers are being met, essentially driving Customer Relationship Management (CRM), as this will allow internal relationships to be strengthened and boost loyalty. From this, churn may be effectively reduced and promote business longevity.

Also with these recommendations, it is important to acknowledge the limitations of our data and how this might impede the accuracy and successfulness of our marketing strategies. The data used in this investigation is severely limited in a number of ways and so above all it is strongly recommended that further analysis is conducted with additional data.

1.0 Situational Overview

The purpose of this report is to investigate the growing increase in customer churn-rate at Telco ABC. **Figure 1** highlights how the average monthly churn-rate has increased from 7.8% in 2000 to 9% 2013 (**Appendix A**). This is a problem for Telco as we are losing customers due to a falling customer retention. Retaining customers is imperative for business success given the positive relationship between loyal customers and profitability (*Helgensen*, *2006*), as current customers are highly valuable assets to business and thus are a detriment to lose. Customer acquisition is also an extremely costly process and so for Telco ABC the best strategy to implement is to reduce their churn-rate and focus on meeting the needs of their current customers (*Reichheld and Kenny*, *1991*).

In order to create an effective communication strategy, data analytics will be employed as this will mean that all decisions will be data driven and increase the chances of success given that the recommendations will be supported by hard evidence. Using data analytics will allow the customer-churn rate to be investigated and unravel critical findings when combined with marketing knowledge.

The structure of this report will then follow with **2.0 Data Quality Report & Methodology** describing the technical process in detail, and leads on to **3.0 Cluster Analysis** which explores our customer base. Following this, **4.0 Findings** will explore the results of the predictive models and will work with the cluster analysis to form our advice to Telco in **5.0 Recommendations.** Finally, the methods and findings are critically evaluated in **6.0 Discussion & Limitations** which also gives recommendations for future customer research.

2.0 Data Quality Report & Methodology

2.1 Data Understanding

In our methodology the CRISP-DM process shall be used and is illustrated in **Figure 2** (**Appendix B)**. All data analysis and prediction models were produced in SAS. The first stage of the CRISP-DM process is to understand the data. The data used here concerns our customer base and the objective is to understand the causes behind customer churn.

With this analysis we will be able to see which variables affect the churn of customers and thus be able to devise new and effective marketing strategies for improved customer retention. **Table 1** below shows the variables contained within the data. The dataset contains 30 variables and 8057 observations.

Table 1: Variables Dictionary

Variables	Type	Description
customerID	Numeric	Contains a unique number for each costumer
Children	Categorical	(true, false) Measures if customer has children.
Credit	Categorical	Customers credit rating. (a, aa, b, c, de, gy, z)
CreditCard	Categorical	(true, false) Measures if customer has credit card.
Custcare	Numeric	Measures the average no. of calls to customer call in the last 6 months.
custcareTotal	Numeric	Measures the total calls to customer call in last 6 months.
custcareLast	Numeric	Measures the call to customer call in the last month.
Directas	Numeric	Measures the number of directory assisted calls made in the last 6 months.
directasLast	Numeric	Measures the number of directory assisted calls made the last month.
Dropvce	Numeric	Measures the dropped calls in the last 6 months.
DropvceLast	Numeric	Measures the dropped calls of the last month.
Income	Numeric	(0 to 9) customer's income
Marry	Categorical	(yes, no, unknown) marital status

Mou	Numeric	Number if minutes last month.
mouTotal	Numeric	The total number of minutes used in last 6 months.
mouChange	Numeric	The presentage of change in minutes.
Occupation	Categorical	The occupation of the customer
Outcalls	Numeric	Measures the number of calls made
Overage	Numeric	Measures the minutes over the customer’s bundle used in this month.
overageMax	Numeric	Max overage
overageMin	Numeric	Min overage
peakOffPeak	Numeric	Measures the total number of peak calls made the last 6 months.
peakOffPeakLast	Numeric	Measures the total number of peak calls made the last month.
Recchrge	Numeric	Measures the recurring bundle charge this month.
RegionType	Categorical	(rural, suburban, town) type of region in which the customer lives
Revenue	Numeric	Revenue from customer last month.
revenueTotal	Numeric	Measures the total revenue in the last 6 months.
revenueChange	Numeric	The percentage of the revenue
Roam	Numeric	Measures the number of roaming events in the time period.
Churn	Categorical	Target Variable(yes, no). Indicates if the customer has churned.

The next step was to observe the variables in the dataset to find any errors and mistakes. This was done with the help of descriptive statistics (**SAS photo 1**) and some visualizations (**Appendix C**) in order to understand and visualise the variables. The following table presents the descriptive statistics of the variables.

Table 2: Descriptive statistics for data understanding

Variable	Missing Values(%)	Min.	Mean	Max.
customerID	0	1000004	1049952	1099988
Children	0	0	0.2404	1
Credit	0	-	-	-
CreditCard	0	0	0.6674	11
Custcare	0	0	1.7250	365
custcareTotal	0	0	8.1164	2253
custcareLast	0	0	1.7302	387
Directas	0	0	0.9069	55
directasLast	0	0	0.8821	52
Dropvce	0	0	6.0154	114
DropvceLast	0	0	5.9603	110
Income	0	0	4.2772	9
Marry	0	-	-	-
Mou	0	0	522.6756	6494.32
mouTotal	0	0	2397.7234	38965.95
mouChange	0	-1	0.6141	967.42
Occupation	74	-	-	-
Outcalls	0	0	159.1477	2716
Overage	0	0	42.0390	4313.14
overageMax	0	0	44.6671	4565.21
overageMin	0	0	39.3981	4098.29

peakOffPeak	0	0	2.0814	79.3300
peakOffPeakLast	0	0	2.0839	79.33
Recchrge	0	0	46.2921	337.98
RegionType	49			
Revenue	0	0	61.7819	577.21
revenueTotal	0	0	279.0563	3463.27
revenueChange	0	-1	0.0431	14.46
Roam	0	0	1.244	742
Churn	0	-	-	-

In the table several values are flagged as being outliers due to their maximum values. These values are highlighted in red and should be treated with caution. Due to this, descriptive statistics were made as mentioned above in order to understand the distribution of each variable (**Appendix C**). The target variable was also checked to understand its distribution.

2.2 Data Preparation

After the preliminary analysis and visualisations, the potential outliers previously identified will now be investigated in further detail. Firstly, the ‘occupation’ variable was removed due to having too many missing values and following this the filtering tool in SAS was used to cut some extreme values from other variables, as shown in **Table 3.**

The selection of the limits as well as any other action in the stage of data preparation was done after studies and tests, to ensure that additional problems or limitations would not follow. Such problems would be for example many new missing values, which would undermine the reliability of the data as well as the predictive power of the machine learning models that followed.

Table 3: Data quality fixing methods

Variable	Data Quality	Fixing Method
CreditCard	1 more category ("unknown")	Delete values with "unknown"
Custcare	Outliers (>30)	Delete values >40
custcareTotal	Outliers (>200)	Delete values >300
custcareLast	Outliers (>30)	Delete values >40
Directas	Outliers (>5)	Delete values >8
directasLast	Outliers (>5)	Delete values >8
Dropvce	Outliers (>30)	Delete values >40
DropvceLast	Outliers (>30)	Delete values >40
Marry	1 more category ("y")	Delete values with "y"
Mou	Outliers (>3000)	Delete values >3000
mouTotal	Outliers (>10000)	Delete values >10000
mouChange	Outliers (>100%)	Delete values >100%
Outcalls	Outliers (>800)	Delete values >800
Overage	Outliers (>850)	Delete values >850
overageMax	Outliers (>850)	Delete values >850
overageMin	Outliers (>850)	Delete values >850
peakOffPeak	Outliers (>20)	Delete values >25
peakOffPeakLast	Outliers (>20)	Outliers >25
Recchrge	Outliers (>150)	Delete values >180
RegionType	1 more category ("unknown")	Delete values with "unknown"
Revenue	Outliers (>250)	Delete values >250
revenueTotal	Outliers (>1000)	Delete values >1000

revenueChange	Outliers (>3%)	Delete values >3%
Roam	Outliers (>100)	Delete values >120

Following these corrections to the data, the total number of entries now sits at 7454, down roughly 600 entries from the original number. Although this is a significant cut to the dataset, it is more important to ensure that any incorrect values are ironed out before further analysis.

After completing the data preparation, the pipeline in SAS was created and contained all of the necessary nodes to complete the project, shown in **Figure 3** below. The 'imputation' node treated any missing values where character variables were the replacement with the distribution of the respective value, while missing numerical variables were replaced with the mean. Following this the tree was divided into two subtrees which had the same parent nodes.

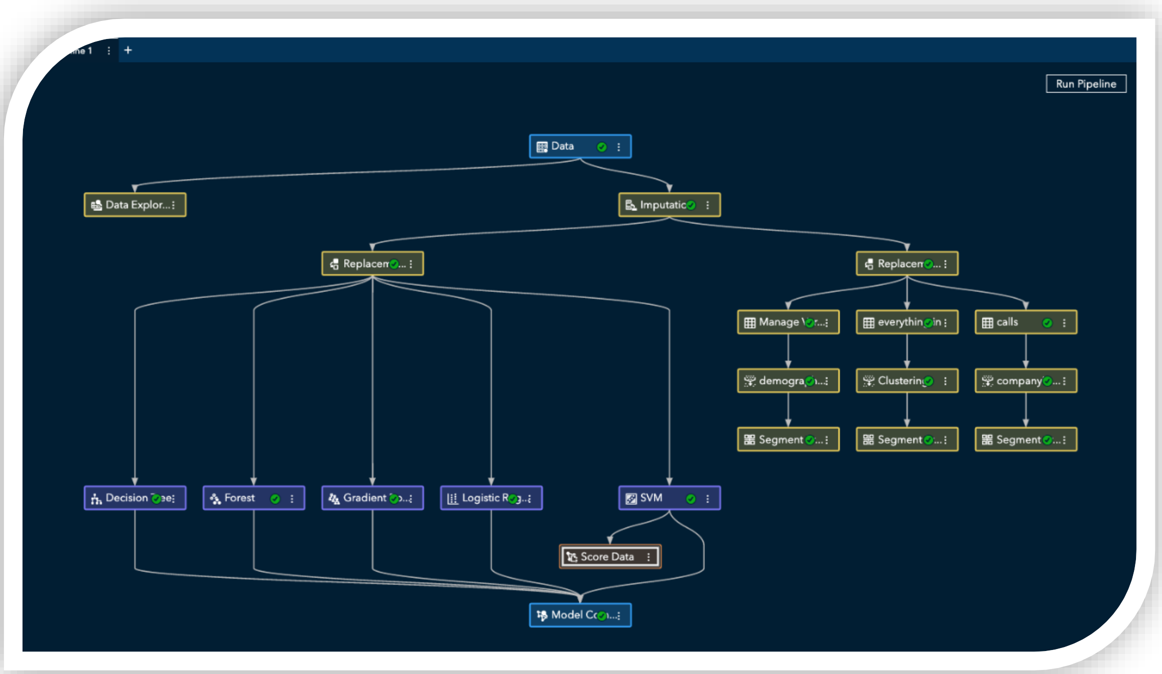


Figure 3: Final Pipeline of the Project

2.3 Clusters

To create the clusters, the variables in the dataset were split into two groups, the demographic characteristics of the customers and another regarding all the other operational and usage variables. In doing so it was assumed and hoped that a proper distribution of clusters would be achieved, although this did not happen due to the poor overall quality of the data. For most features, the vast majority of the data is highly concentrated within small, specific boundaries and often at a value of zero. **Table 4** below contains the details and techniques used to create these 2 types of clusters.

Table 4 - Clustering Strategy

Cluster type	Nodes	Variable selection	Explanation
Demographics	manage variables: keep demographics clustering: k-means = 4	RegionType, children, credit, creditCard, income, marry	Trying to explore the connections between the demographic variables. Understand the connections to create some marketing strategies.
Everything in default	Manage variables: all Clustering: SAS default selection	The whole dataset	Only produced for testing and practice.
Calls	Manage variables: keep everything except demographics Clustering: k-means = 3	Everything except demographics	Trying to explore the connections between variables with telecommunications values.

2.4 Modelling

The next stage of our methodology is to produce the machine learning predictive models. Five models were developed to predict customer churn and learn about its causes. The variety of models produced will make an insightful comparison possible and ensure that the best prediction is achieved. **Table 5** below outlines the key features of each model. The data is split into a training set, validation set and test set in a ratio of 60-30-10 to produce and test these models.

Table 5: Machine Learning Models

Model	Tuning Method	Selected Methods and Tools
Decision Tree	Cross Validation(k=5)	Grow criterion: Information gain ration, Gini, Chi-square, Variance, F-test Search method: Bayesian
Random Forest	Cross validation(k=5)	Grow criterion: Gini, F-test Search method: Bayesian
Gradient Boosted Tree	Cross validation(k=5)	Search method: Bayesian Set 100 trees
Logistic Regression	Without selection method (lasso, stepwise, forward)	Logit method with Newton-Raphson optimization technique
Support Vector Machine (SVM)	Cross validation(k=5)	Kernel polynomial(2nd degree) Search method: Bayesian

Once the models have been produced, they are evaluated using the model comparison node in SAS. The models are scored on their prediction accuracy using a root mean absolute error and a cut-off point of 0.5, meaning that customers scoring above 0.5 will be predicted as likely to churn, while those below will be predicted to be loyal.

2.5 Scoring the Data

The final stage of the methodology is to perform further checks on the models produced. The 'score data' node was used for this and provides the us with the opportunity to evaluate the performance of our models with different data. To do this, new data was uploaded to SAS which contained the same variables but at different values. This node was applied to the best performing model (SVM) and the results were observed. The findings and graphs created from the scoring data will be analysed in **Findings.**

From the data, the following graph (**Figure 4**) was created which shows the distribution of churn predictions (0-no, 1-yes) in each category of churn predictions. In more detail, this graph allows us to see what each category consists of and is used with the supplementary graphs below (**Figures 5,6**) to understand more about the customers' causes of churn.

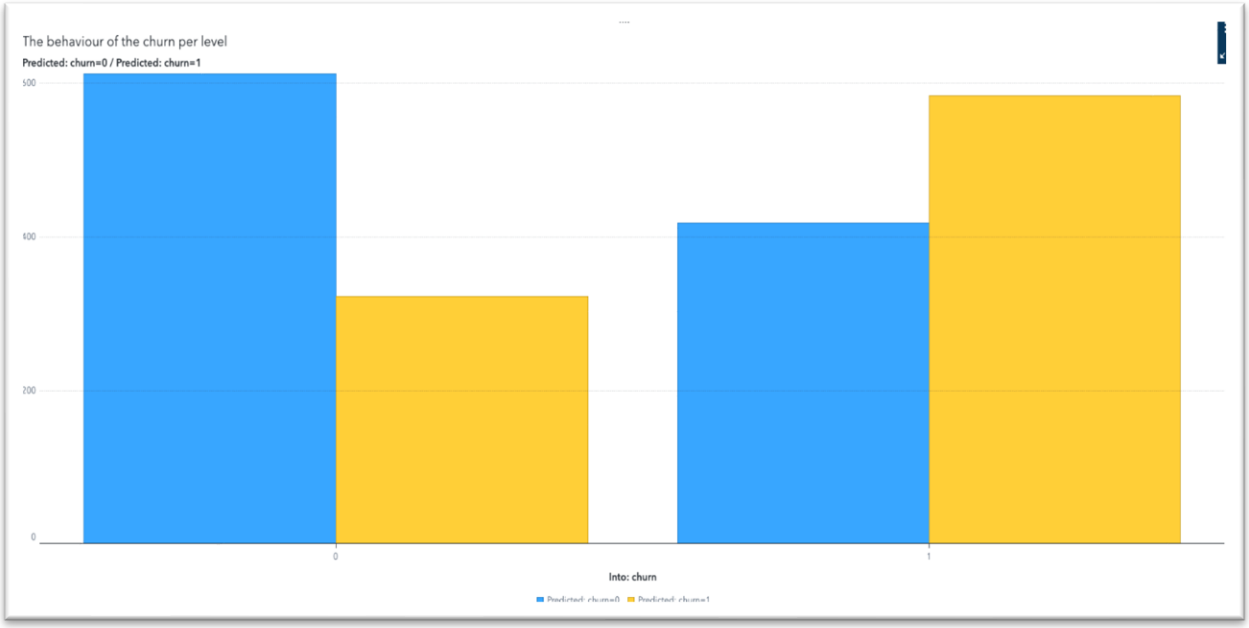


Figure 4: Predicted churns distribution per churn class

The variable 'EM_EVENTPROBABILITY' was also checked in **Figure 7**, which contains the probability that the content of all the data of each client will occur. This means that if the probability is greater than 0.5 it is likely that all the details of a customer are valid. This gives information about the combination of customer characteristics.

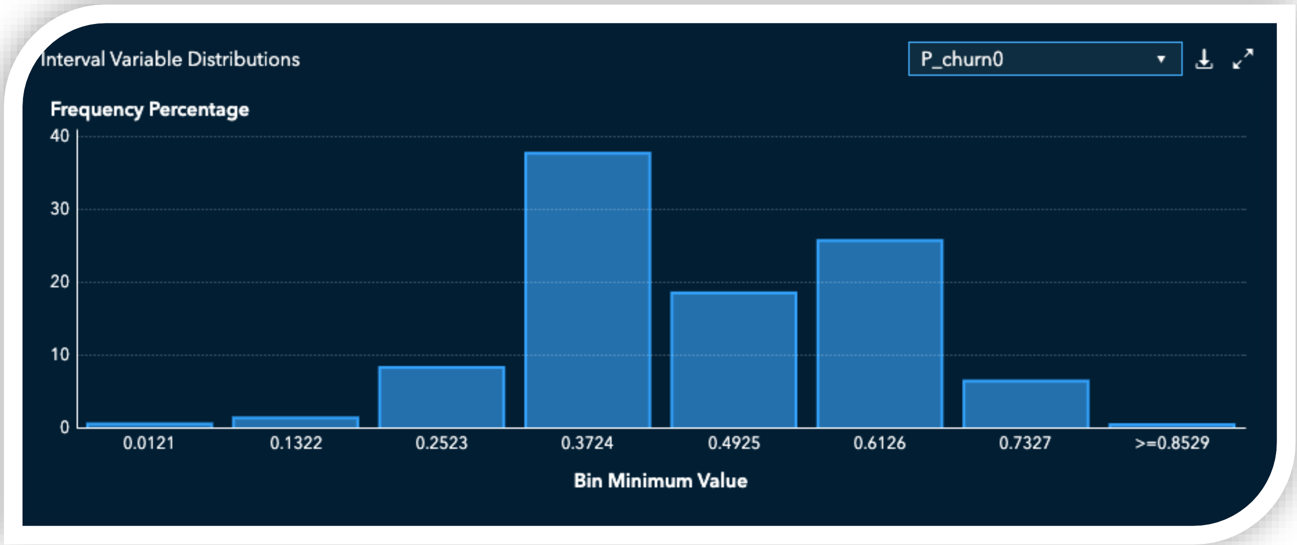


Figure 5: Predicted Churn0 distribution (source: SAS Project)

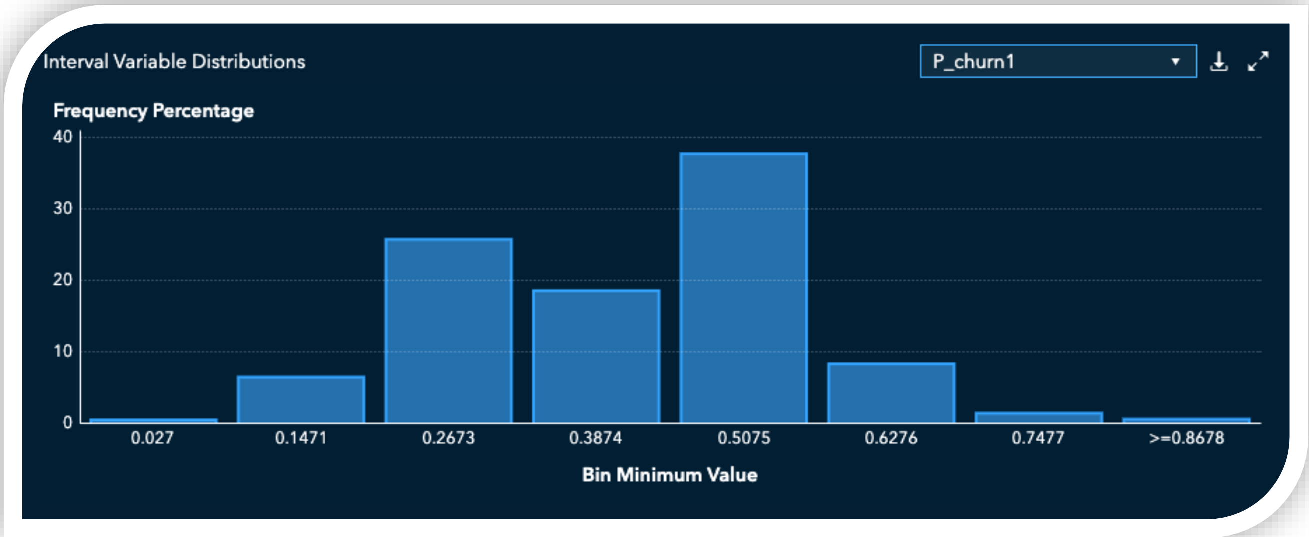


Figure 6: Predicted Churn1 distribution (source:SAS Project)

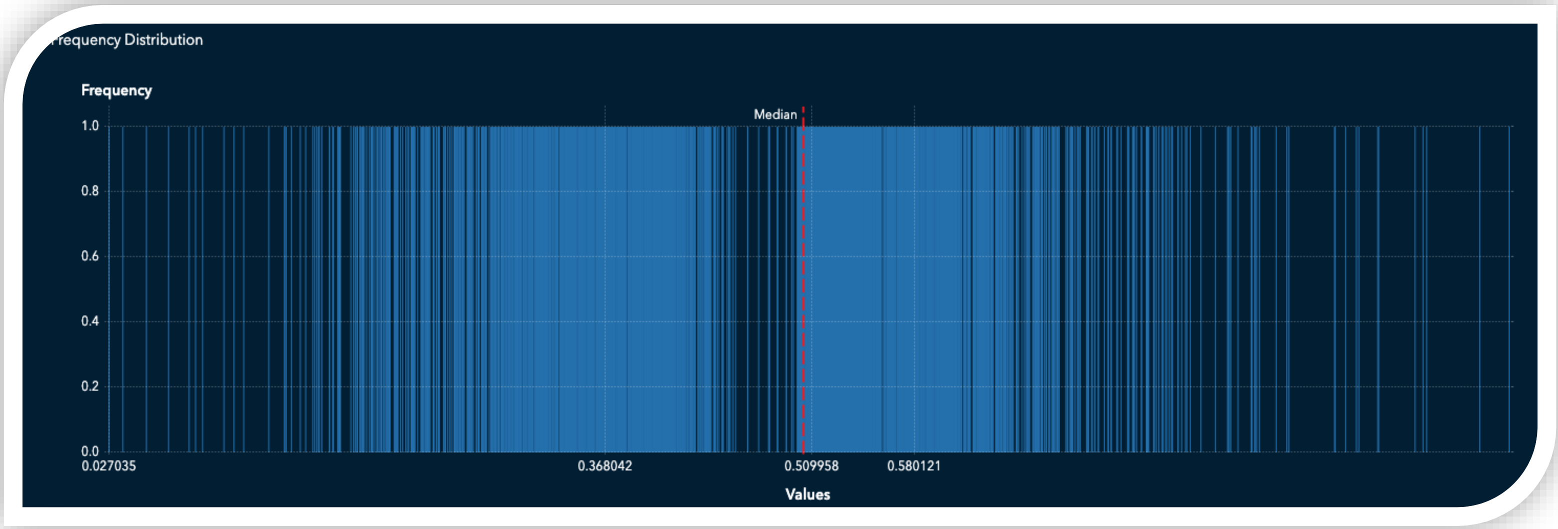


Figure 7: EM_EVENTPROBABILITY distribution

3.0 Cluster Analysis

In CRM, a clear understanding of customer categories is the basis for differentiated management of each customer group. Within Telco, customers of different customer groups show different levels of consumption and require different groups to be treated separately. Therefore, two types of clusters are created through the k-means algorithm, with one cluster considering demographic features and the other exploring usage variables.

Cluster 1 Segment 1

The consumers within this segment are living mostly in suburban areas without any children. Nearly 20% of them have low incomes, or even no income. Among customers with known marital status, more than half of them are unmarried. Then approximately 56% of clients have good credit scores, being A or AA. Only a small proportion have their own credit cards.

Cluster 1 Segment 2

The majority of customers in segment two are unmarried with a high proportion having no children. More than 55% of them live in towns, rather than rural and suburban environments. Nearly 66% of clients in this segment have incomes above level 5 on the scale used in this data and most of these customers have credit cards.

Cluster 1 Segment 3

Nearly half of this customer segment is married, but without children. Around 81% of them have high levels of income, mainly levels 6 – 9 and the majority have good credit scores of A or AA. 85% of these customers live in suburban areas and all have credit cards. These customers have the means for higher levels of service and catering towards them could be more lucrative than other groups.

Cluster 1 Segment 4

The marital status of this cluster is predominately unknown and the vast majority do not have children. Customers within this segment typically have a B credit score and very few have a credit card. Additionally, most customers have lower incomes, equating to below 5 on our scale while as many as 65% are listed as having no income whatsoever. 60% of this group lives in suburban areas and this is likely to be a region of relatively lower rent or house price due to the income bracket of this group. It may include council estates.

Cluster 1 Segment 5

Similar results to segment 3 are observed with this group. Most customers live in suburban areas, followed by those living in towns with the least number living in rural areas, unlike segment 3. Most customers have no children and an unknown marital status with high levels of credit score and income. Combining with the key characteristics of the above four clusters, we confirm that customers who have high levels of income and credit will have their own credit cards.

Cluster 2 Segment 1

Unlike the rest of this cluster, this segment primarily consists of married customers with children and high incomes. More than 50% of them had made directory assisted calls in the last month, and a similar proportion had done so in the last six months. The majority of total minutes for customers is greater than 208 minutes and approximately 75% of consumers have made individual calls lasting more than 73.3 minutes. Given these affluent demographics and high usage statistics, these customers should be valued highly.

Cluster 2 Segment 2

The majority of this segment do not have children and have an unknown marital status. Similarly to segment 1, most of them have high incomes above 6 on our scale. 81% of this group did not make directory assisted calls in the last month and few did so in the last six months. In fact, 49% of this group didn't make any calls at all in the previous month and so we may infer that this group has another channel as their primary communication method, likely to be through a social media platform. More than 85% of customer revenue is 9.9-66.3, indicating that the customer base is of medium to low value. This segment clearly shows an unsurprising relationship between the numbers of calls, call minutes and revenue. When customers didn't make a phone call in the last month, the revenue was less than 28.7.

Cluster 2 Segment 3

Similar to segment 2, the marital status of most customers in this segment is unknown and the number of directory assisted calls made in the last 6 months by this group are mostly zero. Most people within this segment made phone calls last month, with 58% of customers having made calls for more than 73 minutes and 65% of them being called for more than 208 minutes. The fact that this group uses significant minutes but does not use directory assistants suggests that this group is younger than others with individuals who find phone number online rather than through a directory.

Cluster 2 Segment 4

Segment 4 reports similar results to segment 3, with approximately 70% of people within this segment never making a directory assisted call in the last six months and having similar usage statistics in regards to minutes and number of calls. Unlike segment 3 however, most people here live in suburban areas with the remaining 40% being split between rural and town areas. This cluster found comparable revenue to segment 3, with more than 70% of them paying 28.7-66.3 for Telco products. From this and other clusters we learn that the location of the customer is largely irrespective of the revenue they generate.

4.0 Findings

4.1 Decision Tree

The five predictive models used were found to have incredible accuracy, but with significant limitations in the insights they produced. The decision tree modelling was capable of correctly classifying customer churn with 98.8% accuracy, meaning that of the 745 customers in the testing set, only nine were incorrectly classified. This low error is made possible by the pruning shown in **Figure 8**, where using an increasing number of leaves leads to a lower overall error. Here, 36 leaves are used and correspond to a misclassification rate of 0.0121. Although this accuracy is very impressive, it is surprising to learn that the model does this with just three variables and could be almost just as accurate with just two of those.

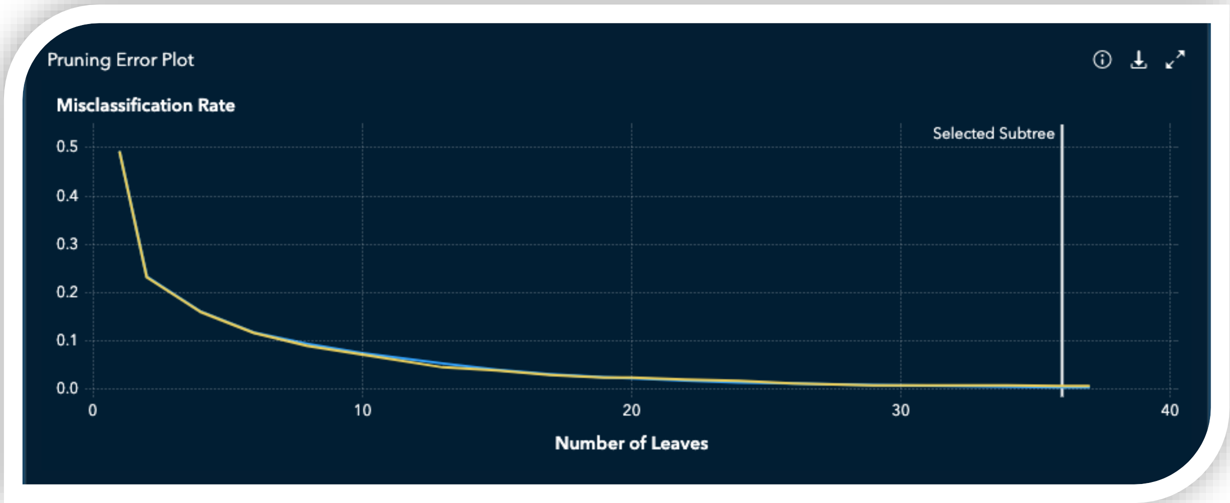


Figure 8: Missclassification rate with increasing number of leaves in the decision tree model

The decision tree model only needs to learn from the variables 'RevenueTotal', 'Revenue' and 'DropVCE'. The total revenue from the customer in the last six months has a negative relationship with churn, as is also observable in **Figure 9**, where we see that there is a pattern of increasing total revenue equating to a decreasing proportion of churned customers. This relationship forms the biggest predictor of the model, and has a relative importance of 719.4, compared to 366.6 for 'Revenue' and just 1.2 for 'DropVCE'. It is both curious and unfortunate that 'RevenueTotal' is around 600 times more important to prediction than DropVCE, and any other variables that are not used would have been even less significant. Despite the model only using three variables, one of which being far less important than the other two and meanwhile the other two being very similar in nature, the model does still hold strong predictive power and informs us of the key link between churn and revenue.

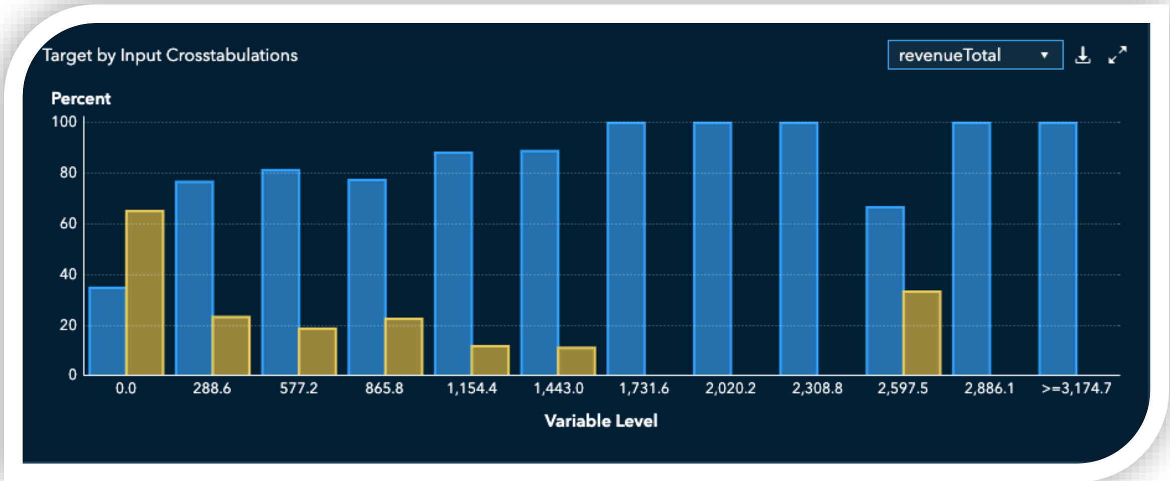


Figure 9: Proportion of customers that churned for each bracket of total revenue.

4.2 Random Forest

The random forest model paints a similar portrait to the decision tree but does find prediction value from additional variables due to its nature as a model that depends on comparing different trees. ‘RevenueTotal’ and ‘Revenue’ once again make up the top two predictors, with a training importance of 600.0 and 349.1 respectively. Following these however, ‘MouTotal’ and ‘Mou’ (minutes used by the customer in the last 6 months/ month) both have a relative importance of around 20, while ‘Recchrge’ (the recurring bundle charge) has an importance of 10.9. These variables and several others of lower importance are used in the 50 trees of the model to predict churn with an accuracy of 98.3%. These additional variables tell us some new links between the data, notably that a higher recurring monthly bundle charge leads to a higher likelihood to churn, and that the minutes used in the last month and six months are weak predictors unless there has been a significant change in usage in the last few months for a customer.

4.3 SVM & Gradient Boosted Tree

The gradient boosted tree method works very similarly to the other models and predicts churn with an accuracy of 98.8% by primarily using the revenue variables and touching upon the minutes used and bundle charge variables. Meanwhile the support vector machine model uses a 2nd degree kernel polynomial to achieve the highest accuracy of 99.7%, using variables in common with the previous models.

4.4 Logistic Regression

Logistic regression is our second most accurate model (although the differences are negligible) with an accuracy of 99.5%. The majority of variables are used and the breakdown of this model tells us about the relationships between some of the variables and churn more clearly than graphical visualisations (**Appendix C**). Firstly, we learn that not only do we have more customers with better credit scores, but these customers are also less likely to churn. This makes logical sense given the nature of credit scores and we can be confident in this finding. We also note that although the impact is small, the number of calls made to a customer (from Telco) does impact their likelihood to churn. Customers who received more calls from Telco churned less than those who received fewer or none. Although this is a great insight into the fact that the calls made to customers are often successful, this relationship has not been tested at high frequencies of calls and so it is not advised to call customers excessively. An additional insight is it is found that there is some link between increasing use of overtime minutes and an increasing likelihood to churn. The fit statistics are given in **Table 6** below.

Table 6: Fit statistics for the logistic regression model

Description	Training	Validation	Testing
-2 Log Likelihood	0.00058376	184.63549	129.06795
AIC (smaller is better)	16.00058	200.63549	145.06795
AICC (smaller is better)	16.03285	200.70015	145.26334
SBC (smaller is better)	67.24531	246.33504	181.98576
Average Square Error	1.3719E-12	0.00270	0.00555

4.5 Scoring Data Analysis

The use of the 'score data' node allowed for further review of the models' accuracy. **Figures 5,6** in the methodology showed that the predicted values for churn have a high concentration around 0.5, which is chosen as the cut off point. The same goes for the regular dataset and suggests that the clients' decision to churn was very marginal. The implication of this is that whether customers are predicted to churn - and if they do in reality – are both incredibly volatile and susceptible to be changed. The good news from this is that it means there is large scope to reduce churn massively, but the bad news is that any mistakes in a Telco marketing strategy have the potential to lose many customers instead.

5.0 Recommendations

5.1 Customer Segments Justification

Segment 1: Families

From our data, our first customer segment identified is the family-oriented group, as the majority of our customers fall in the relationship status of married and then within that most also have children. In terms of demographics, most of our customers have good credit scores of either A or AA and live in suburban areas. This data paints the picture of a large portion of our customers being in a financially secure family and serve as a good group for us to target our marketing towards.

Segment 2: Students

The second observed segment was found to be students, as there was a good portion of customers who were single in the data set without children and more volatile credit scores. This identification of students is further supported by the pattern of lower incomes and often no income whatsoever. Additionally, the number of these customers that are in towns are higher in proportion than other groups, which suggests that these may be students living in university towns, which are typically much more student populated than rural areas or suburban areas designed for families.

The Target Customer

For a good marketing strategy is it imperative that we meet the needs of customers therefore we must identify who are target is in order to a provide a sufficient strategy. In our recommendations, families will be our target customer as they were identified as the biggest proportion in our customer base and additionally gives us a foot in the door to sell our products to all members of a family that have a customer with us.

In our marketing strategy we aim to focus on our customer-relationship management (CRM) as we want to ensure our customer needs are met and that we are creating strong internal relationships. This will ensure business succession by boosting customer loyalty and help solve the churn issues within our business (*Debnath et al., 2016*). To drive CRM, this will be how we market and communicate our services with our customers – focusing on making sure our current customers (and particularly families) are satisfied over trying to gain new customers.

5.2 Our Recommendations

We will offer a 'family bundle', whereby there can be multiple contracts under one account. This means that parents can keep their account but now add the members of their household to their bill. This simplifies their mobile contracts and keeps everyone in their household with our services rather than any competitors. Within this we would offer a discount on each additional phone added, the specifics of which must be discussed. This would encourage additional purchases for families given they are rewarded for having a larger household rather than penalised. Furthermore, family members are considered to be the most trusted recommendation source by 70% (*Jankowski et al., 2014*) and so individuals are highly likely to use our services if their family members have had a positive experience.

Secondly, there could also be the ability for families to share data within their household, to prevent cellular data wastage. This would be particularly beneficial for families with teenagers and older children, whereby they could use any wasted data perhaps from their parents. Given the assumption that younger dependents will use more data, due to their relationship with high mobile internet usage (*Statista, 2021*). This would have to be a tactical decision however, as it would potentially cause us to miss out on significant revenue.

The introduction of our new offering will ensure that we are meeting the needs of families as we are accommodating for their larger sized household as well as providing the ability to share. As to market our services, we want to position ourselves in a way that we are about relationships, value and community, as these are values that will resonate with our target customer.

Value will then be emphasised in how our family bundle will save customers money and give them more control over their children's spending. This package also meets their needs of stability, given that everyone can be under one account, so any risk is minimised in that everyone's phone can be controlled under one name. It also meets families' desire to be a community and one unit. The sharing functionality also furthers this need, re-emphasising value.

To deploy this package our main marketing channel to contact our customers will be through telephone calls as based on our data findings this was an effective strategy given its relation to reducing churn. Therefore, this a preferred medium and will thus be used to meet the needs of our customers. Furthermore, our marketing messages and style will be very family orientated and will be focused on supporting one in other to install a sense of community, as we want to create a brand that is about togetherness and customer relationships. This will ensure that our values and mission align with the needs and values of our family customers. Creating a system based on community and value also aids trust and increasing purchase intention and loyalty (*Hur et al., 2011*). We also want to ensure we are listening and responding to their needs, shifting away from any perception of us as profit-driven. This move is done in an attempt to improve loyalty.

6.0 Discussion & Limitations

The incredible predictive accuracy of the models produced and the insights drawn from our clustering are overshadowed by the poor quality of data they came from. Although only 600 out of 8000 customer entries were removed entirely, so many of the remining customers have many variables unknown. This is especially true for the occupation variable which was removed entirely due to only having values for less than 30% of our customers. Had more information on each customers occupation been acquired, further insights may have been produced. This might include learning that the location people work in is also important in their decision for a mobile contract. Or we might learn that a particular industry is more receptive to our marketing and focus more resources there.

What is even more detrimental than the missing values, however, is the absence of so many key demographic variables. The dataset does not include age or gender and this combined with the lack of occupation data makes it very difficult to confidently categorise a customer. Missing an age variables does not just limit our capabilities of discovering what ages we are most and least successful with, it causes mass uncertainty in our dataset by removing our ability to perform logical checks. For example, a surprisingly large proportion of our customers are listed as having no income. This may be due to incorrect data, having children in the dataset, having students in the dataset or due to having adults who do not have a job but live with a partner who does. Having no way to infer the age of our customers makes it near impossible for us to confidently conclude which of these it is, and this is further complicated by the categorisation of the income feature, where zero may actually include a small amount of income but still be the smallest bracket. Similarly, not knowing what an income of five or nine is equivalent to numerically makes it hard to assess how wealthy our customers actually are and restricts us to speak about such only relatively.

Of the variables that were used, the distribution of values was often very poor and lead to many variables being largely insignificant when compared to others. **Figure 10** below illustrates how the revenue variables were overwhelmingly more impactful than any others. Moreover, the lack of knowledge of the particular area Telco is operating complicates the insights from several other variables. It is unknown what currency the revenue is recorded in and the regions of where customers live are poorly defined. Although we know that the majority of our customers come from suburban areas, we do not know if this is just reflective of the distribution of people in the region as a whole, and so this data should be accompanied by local census data in future analysis.

Although this investigation has successfully developed a marketing strategy and drawn key insights from the data, significant further research is strongly advised, and it is vital that such is done with better data collection.

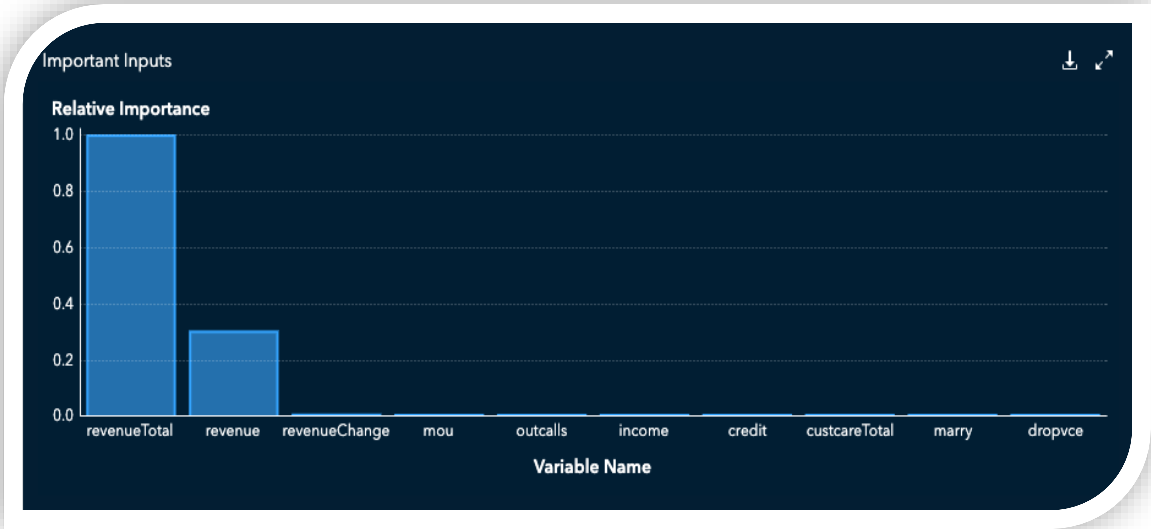


Figure 10: The relative importance of the variables in the dataset.

7.0 Bibliography

Debnath, R., Datta, B. and Mukhopadhyay, S (2016) ‘Customer relationship management theory and research in the new millennium: Directions for future research’, Journal of Relationship Marketing, 15(4), pp.299-325 [Online] Available at: https://www.tandfonline.com/doi/pdf/10.1080/15332667.2016.1209053?casa_token=Jf6gR_M_yfYAAAAA:LRSPisOHJ13i28kl7KuBERK-4kGlUFg45W7mjcACFS6vrkFAobifj7dU4_hhEO9KVhz2QBwsAQl4Bw (Accessed 1st May 2021)

Helgesen, Ø (2006) 'Are loyal customers profitable? Customer satisfaction, customer (action) loyalty and customer profitability at the individual level', Journal of Marketing Management, 22(3-4), pp.245-266 [Online] Available at: https://www.tandfonline.com/doi/pdf/10.1362/026725706776861226 (Accessed: 20th April 2021)

Hur, W.M., Ahn, K.H. and Kim, M. (2011) ‘Building brand loyalty through managing brand community commitment’, Management Decision. [Online] Available at: https://www.emerald.com/insight/content/doi/10.1108/00251741111151217/full/html?casa_token=s9D3OVS_Bj0AAAAA:smh-1q5Joeu-EnCZXD8mrMq5afVjY5k-1vlkZZ3tqpqm3GHwqJsBRTmi3gzQ5BjayqEnwSlptuJK67FOagkCV036UTZcvCr9ELqB7K5O4_h8A6Napu02(Accessed 1st May 2021)

Jankowski, P (2014) 'What Brands Need To Know About Families--We're All Dysfunctional', Forbes, 20th November, [Online] Available at: https://www.forbes.com/sites/pauljankowski/2014/11/20/what-brands-need-to-know-about-families-were-all-dysfunctional/?sh=749017b41d46(Accessed: 1st May 2021).

Reichheld, F.F. & Kenny, D.W. (1991) ‘The Hidden Advantages of Customer Retention’, Journal of Retail Banking, vol. 12, no. 4, pp. 19 [Online] Available at: https://www.proquest.com/docview/214539065?accountid=13374 (Accessed 30th April 2021)

Rokach, L. Maimon , O.Z. (2008) Data Mining with Decision Trees Theory and Applications, Singapore: World Scientific.

Statista (2021) Share of individuals who accessed the internet via a mobile phone in Great Britain in 2019, by age and gender, Available at: https://www.statista.com/statistics/275985/mobile-internet-penetration-in-great-britain-by-age-and-gender/ (Accessed: 1st May 2021).

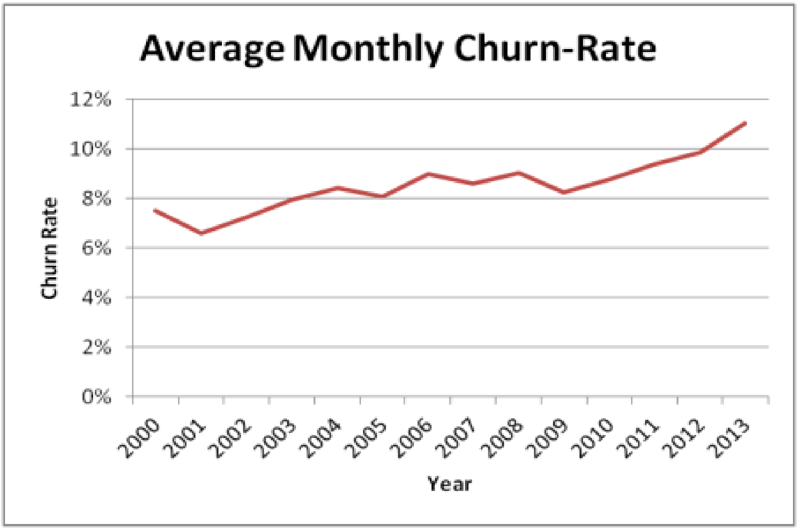
Timonina-Farkas, A., Katsifou, A. and Seifert, R (2020) ‘Product assortment and space allocation strategies to attract loyal and non-loyal customers’, European Journal of Operational Research, 285(3), pp.1058-1076.

Winer, R., (2001) ‘A Framework for Customer Relationship Management’, California Management Review, 43(4), pp.89-105.

8.0 Appendices

Appendix A – Historical Telco Churn

Figure 1: Average Monthly Churn-Rate



Appendix B – CRISP-DM Process

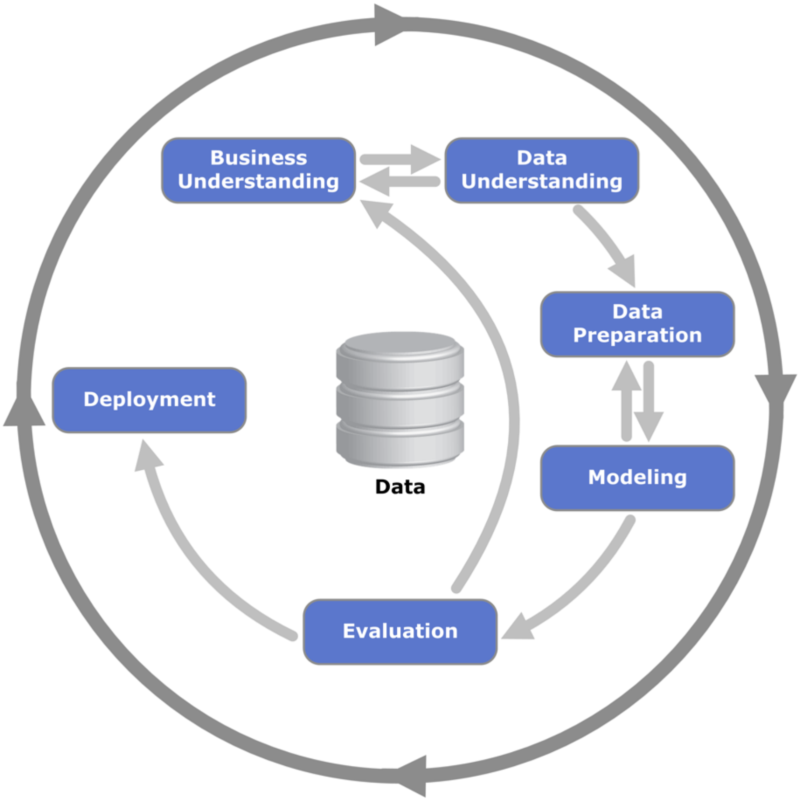
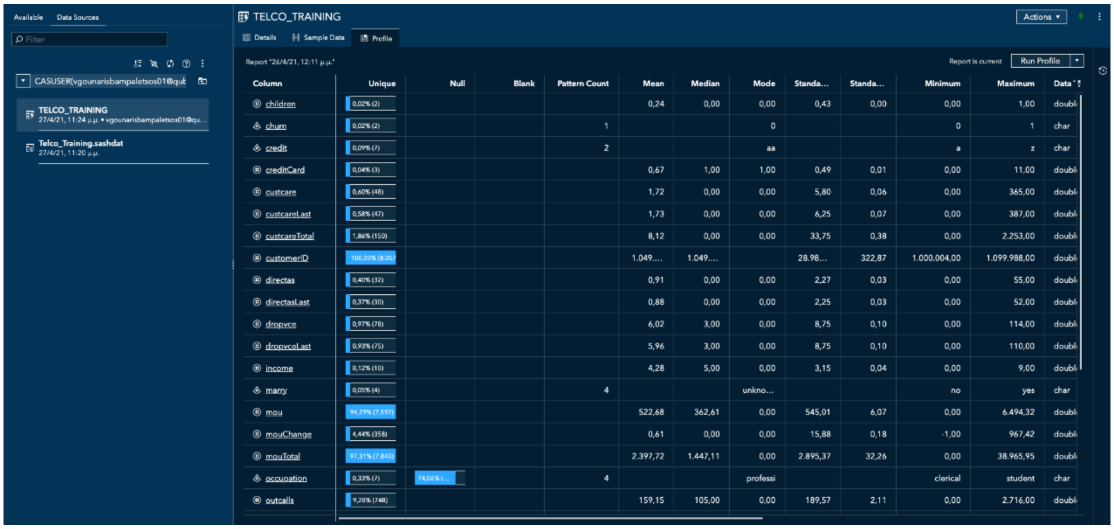


Figure 2 - Crisp-dm process diagram (source: https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)

Appendix C – Supplementary Graphs

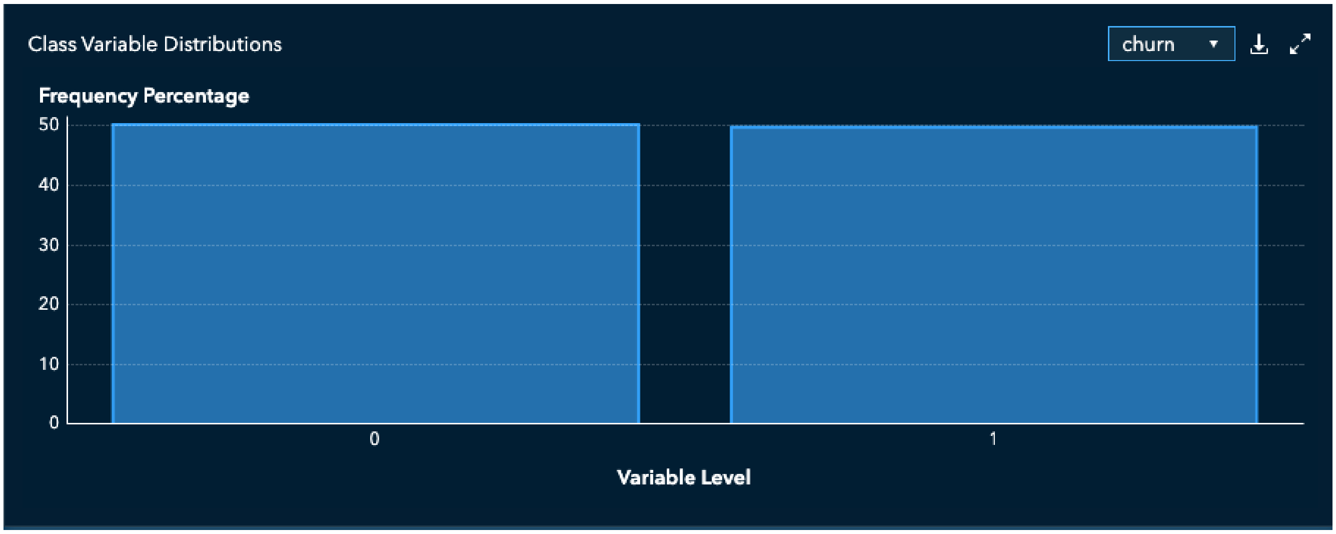
SAS – Data Understanding/Summary Statistics



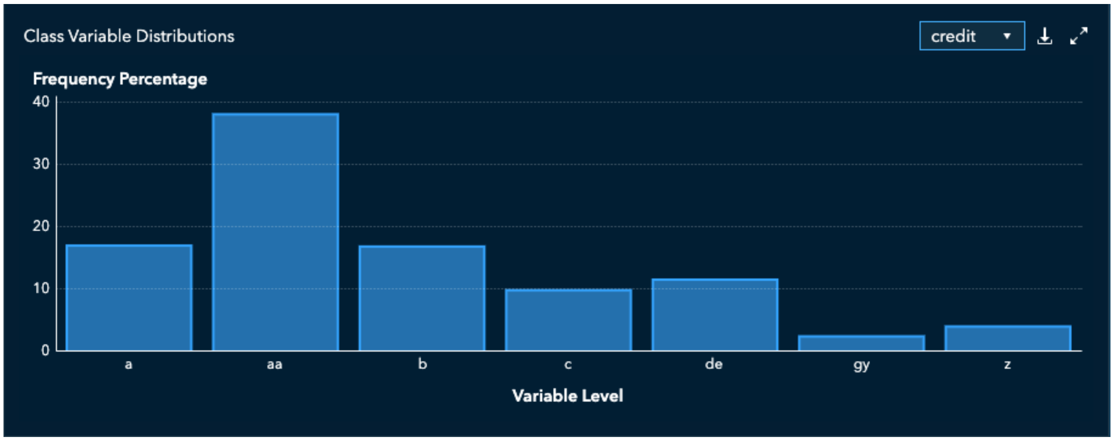
Report "26/4/21, 12:11 μ.μ."

Column	Unique	Null	Blank	Pattern Count	Mean	Median	Mode	Standa...	Standa...	Minimum	Maximum	Data
children	0,02% (2)				0,24	0,00	0,00	0,43	0,00	0,00	1,00	doubl
churn	0,02% (2)			1			0			0	1	char
credit	0,09% (7)			2			aa			a	z	char
creditCard	0,04% (3)				0,67	1,00	1,00	0,49	0,01	0,00	11,00	doubl
custcare	0,60% (48)				1,72	0,00	0,00	5,80	0,06	0,00	365,00	doubl
custcareLast	0,58% (47)				1,73	0,00	0,00	6,25	0,07	0,00	387,00	doubl
custcareTotal	1,86% (150)				8,12	0,00	0,00	33,75	0,38	0,00	2.253,00	doubl
customerID	100,00% (8.057				1.049....	1.049....		28.98...	322,87	1.000.004,00	1.099.988,00	doubl
directas	0,40% (32)				0,91	0,00	0,00	2,27	0,03	0,00	55,00	doubl
directasLast	0,37% (30)				0,88	0,00	0,00	2,25	0,03	0,00	52,00	doubl
dropvce	0,97% (78)				6,02	3,00	0,00	8,75	0,10	0,00	114,00	doubl
dropvceLast	0,93% (75)				5,96	3,00	0,00	8,75	0,10	0,00	110,00	doubl
income	0,12% (10)				4,28	5,00	0,00	3,15	0,04	0,00	9,00	doubl
marry	0,05% (4)			4			unkno...			no	yes	char
mou	94,29% (7.597)				522,68	362,61	0,00	545,01	6,07	0,00	6.494,32	doubl
mouChange	4,44% (358)				0,61	0,00	0,00	15,88	0,18	-1,00	967,42	doubl
mouTotal	97,31% (7.840)				2.397,72	1.447,11	0,00	2.895,37	32,26	0,00	38.965,95	doubl
occupation	0,33% (7)	74,06% (...		4			professi			clerical	student	char
outcalls	9,28% (748)				159,15	105,00	0,00	189,57	2,11	0,00	2.716,00	doubl

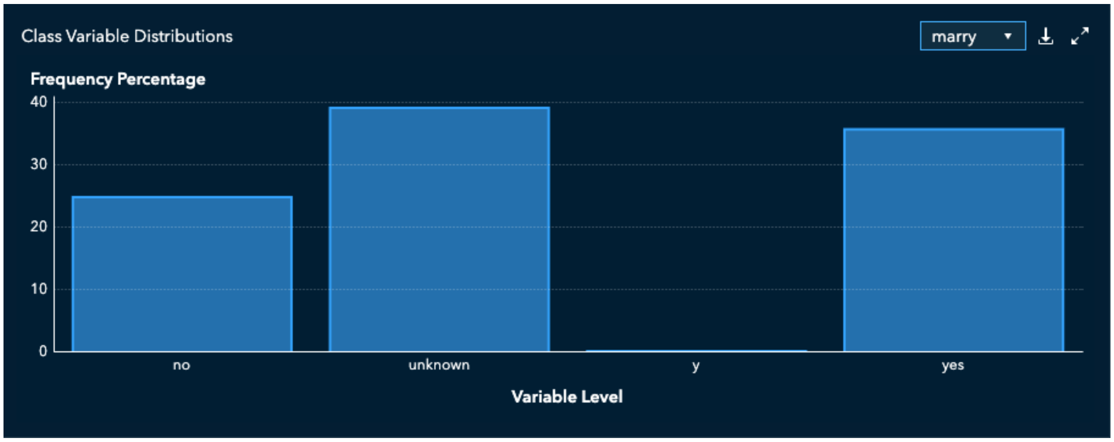
SAS photo 1: Summary statistics



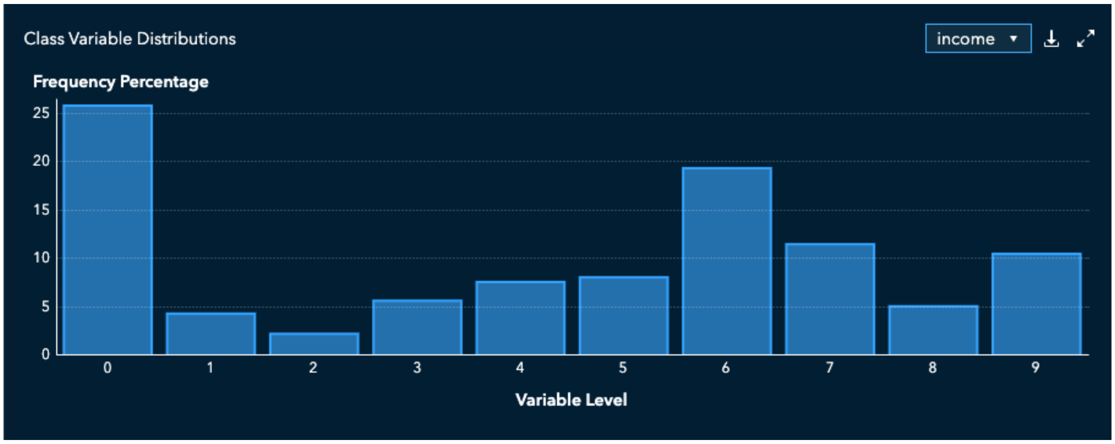
SAS photo 2: Churn distribution



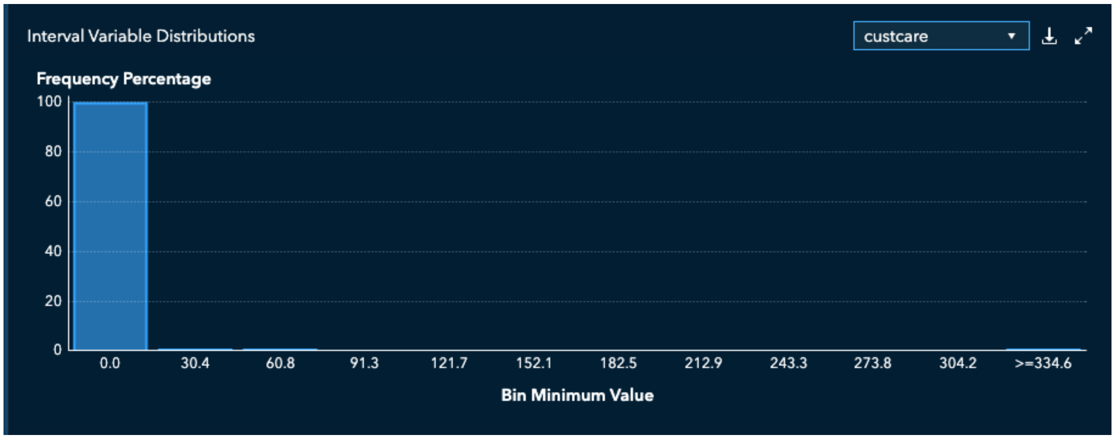
SAS photo 3: Credit distribution



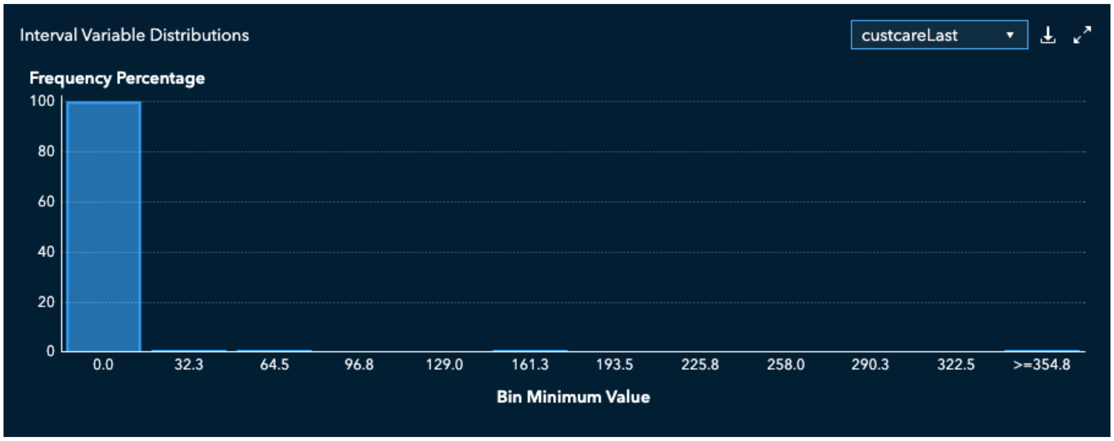
SAS photo 4: marital status distribution



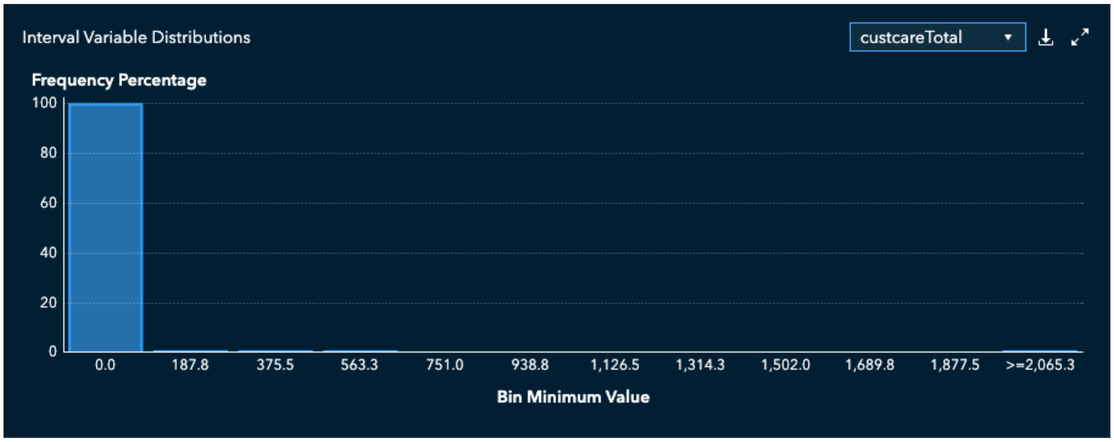
SAS photo 5: income distribution



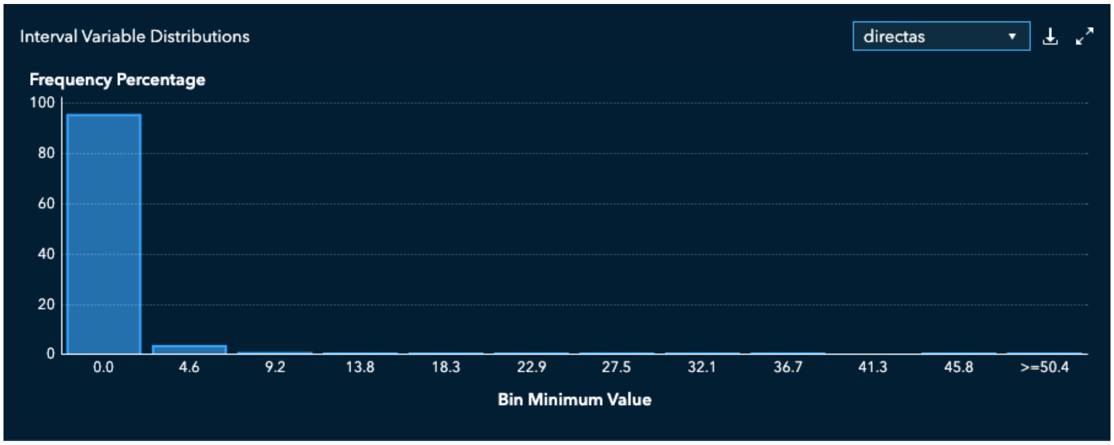
SAS photo 7: custcare distribution



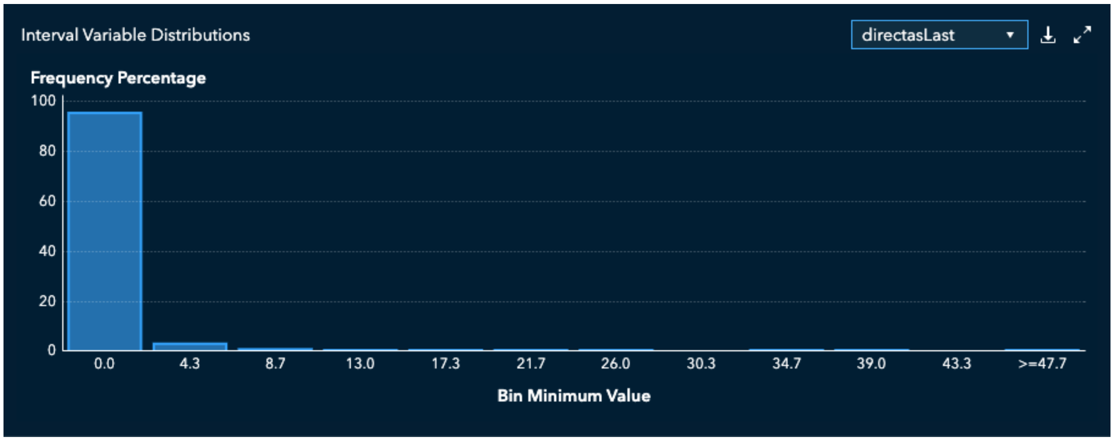
SAS photo 8: CustcareLast distribution



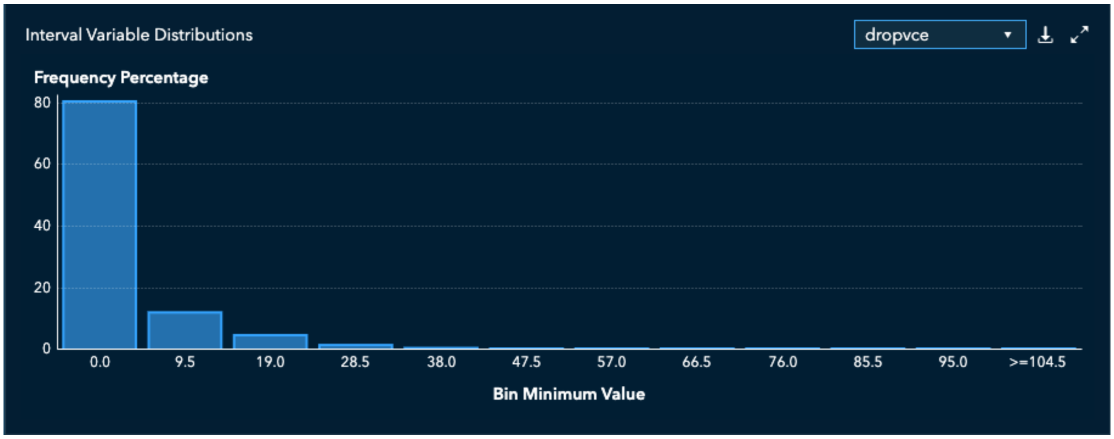
SAS photo 9: custcareTotal distribution



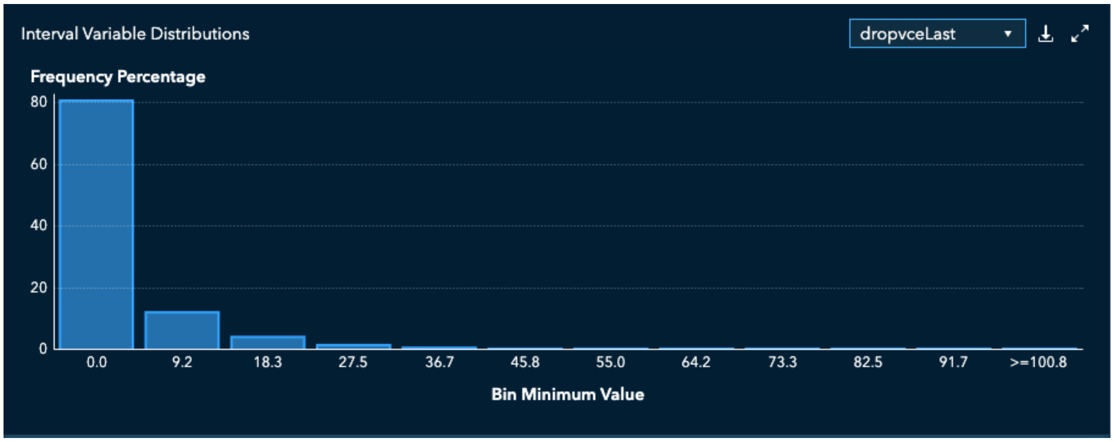
SAS photo 10: directcas distribution



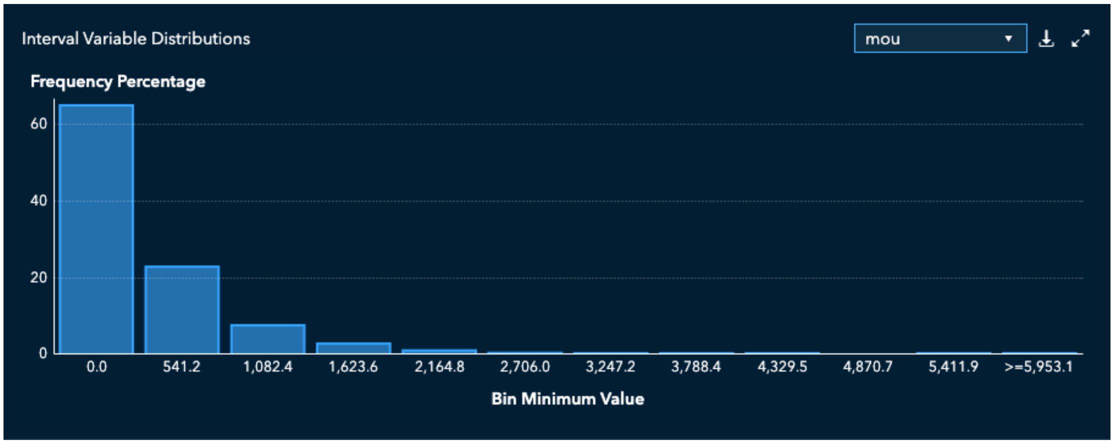
SAS photo 11: directcasLast distributio



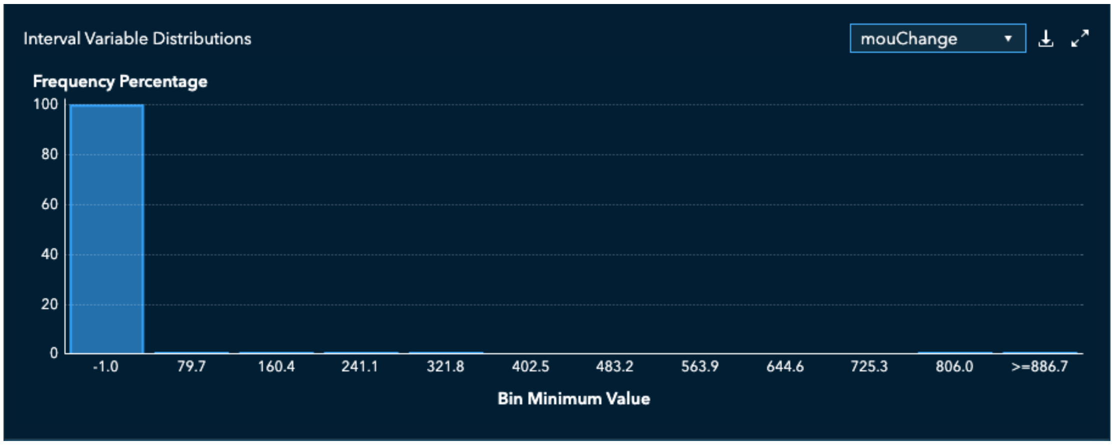
SAS photo 12: dropvce distribution



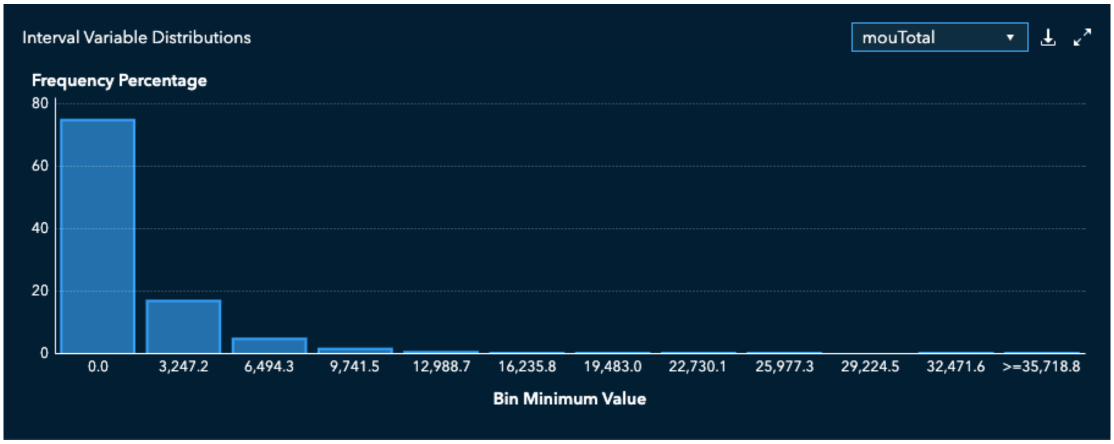
SAS photo 13: dropvceLast distribution



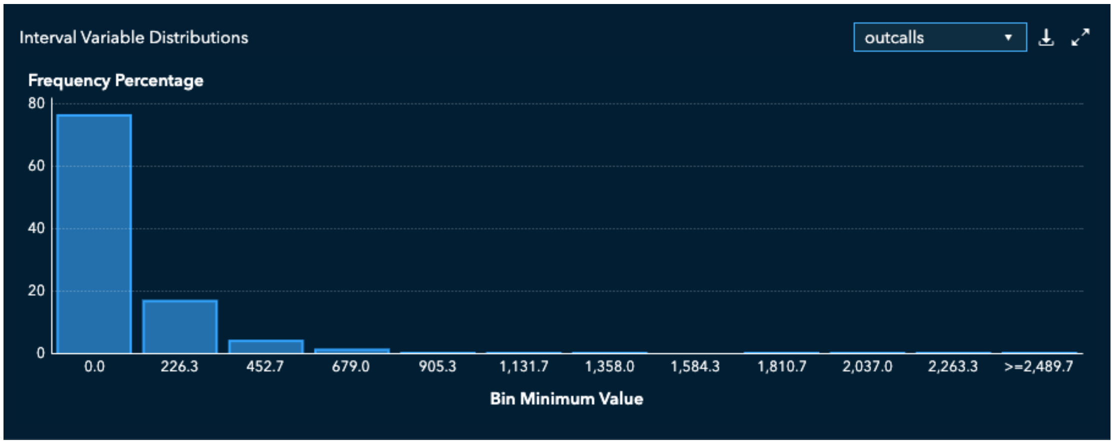
SAS photo 14: mou distribution



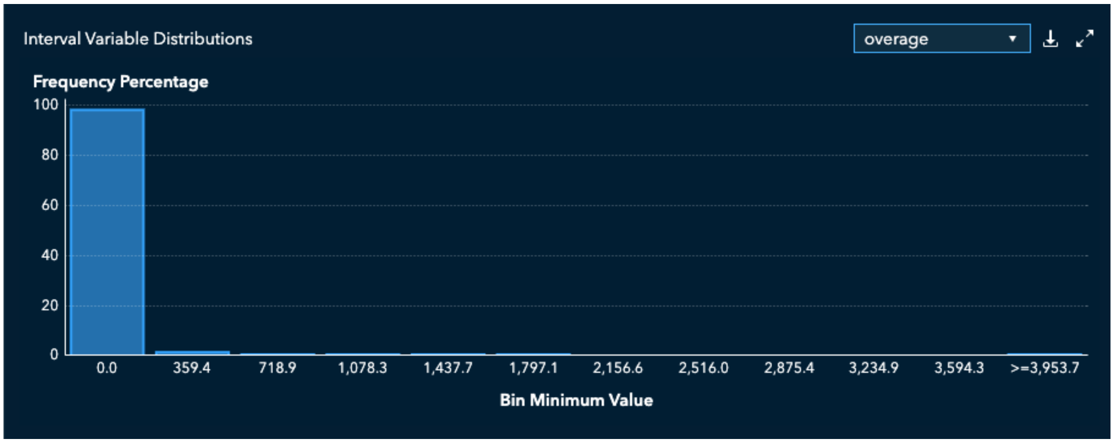
SAS photo 15: mouChange distribution



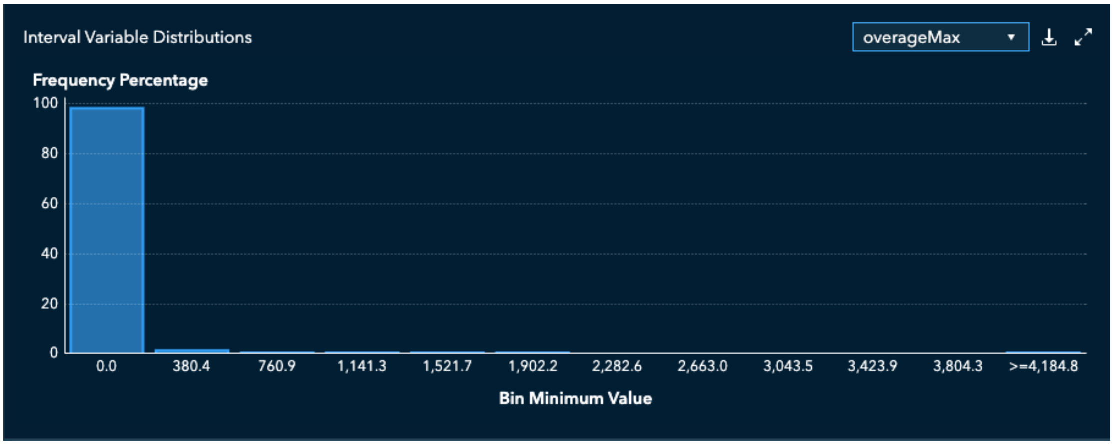
SAS photo 16: mouTotal distribution



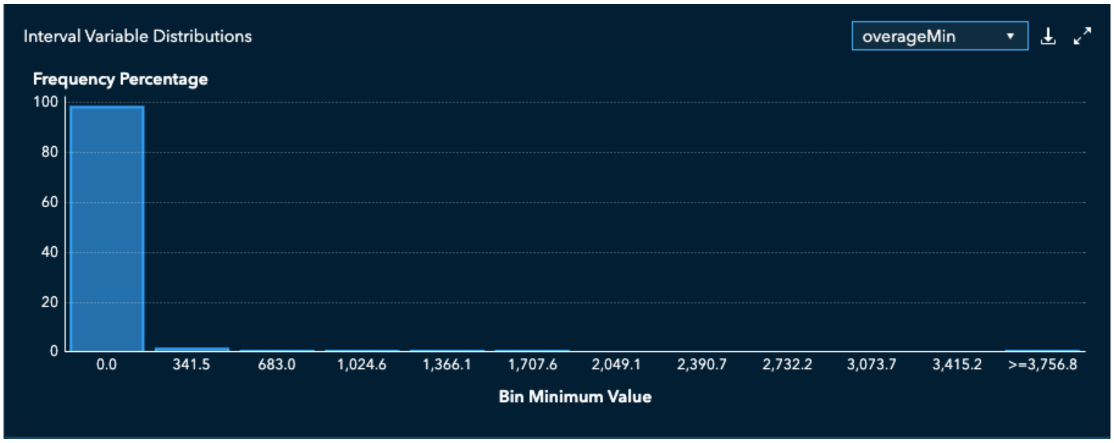
SAS photo 17: outcalls distribution



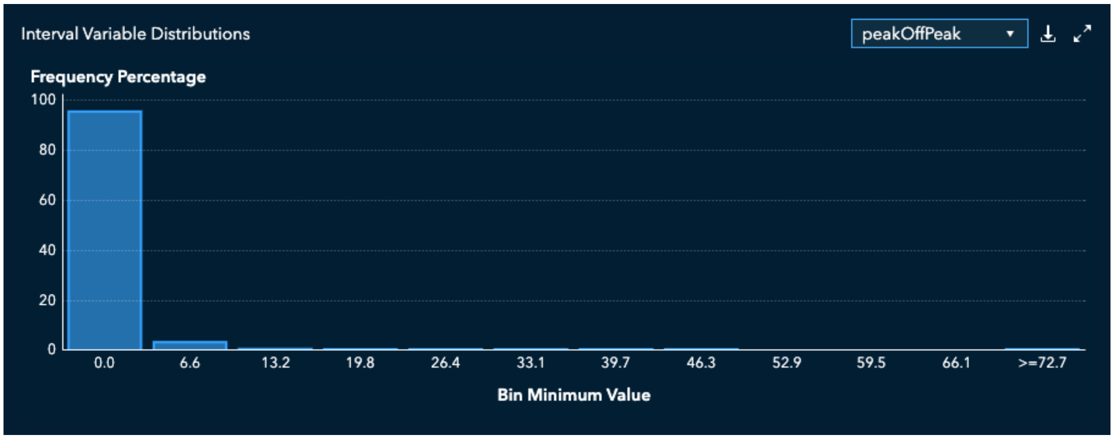
SAS photo 18: overage distribution



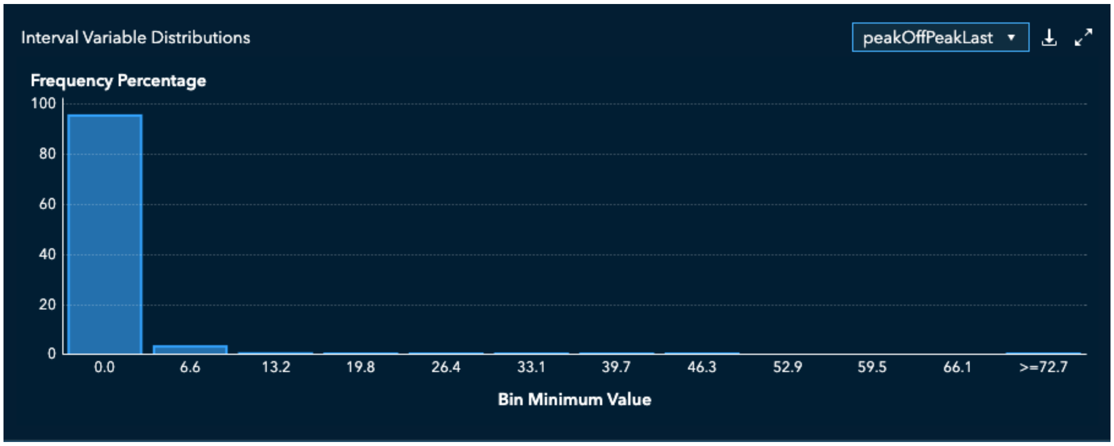
SAS photo 19: overageMax distribution



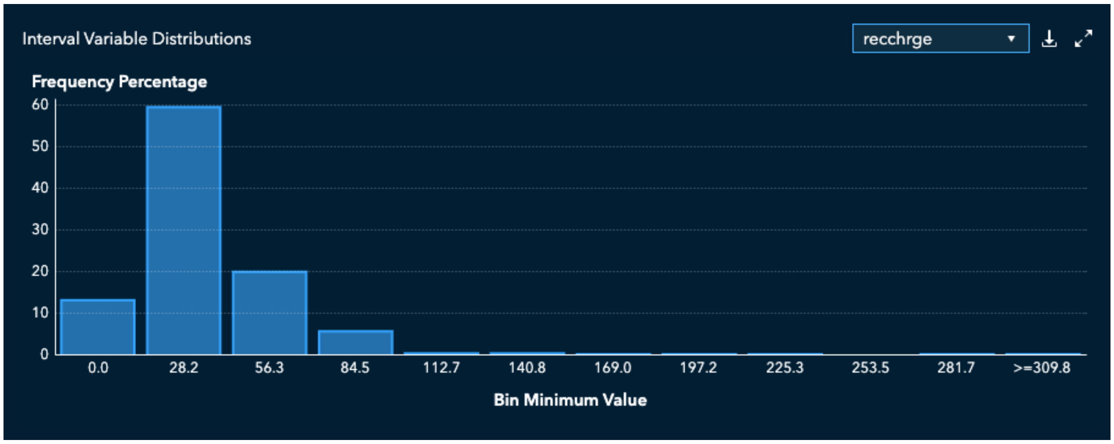
SAS photo 20: overageMin distribution



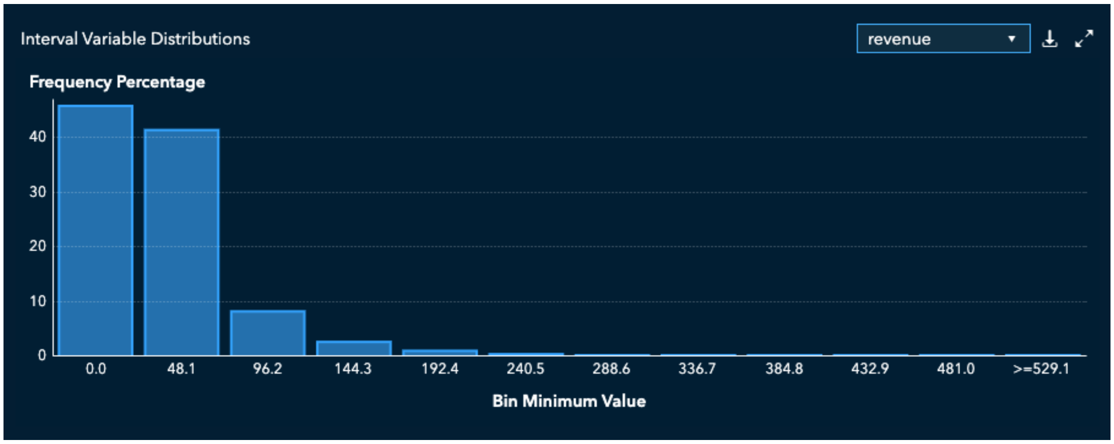
SAS photo 21: peakOffPeak distribution



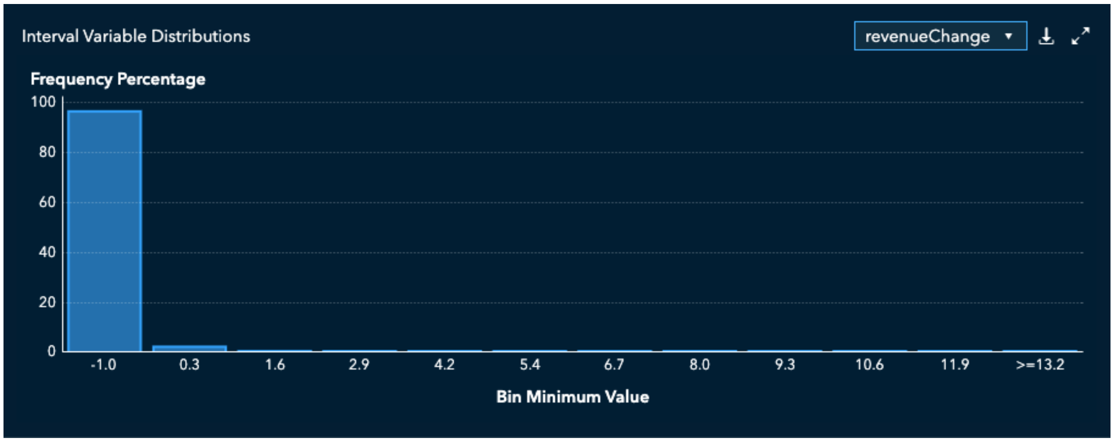
SAS photo 22: peakOffPeakLast distribution



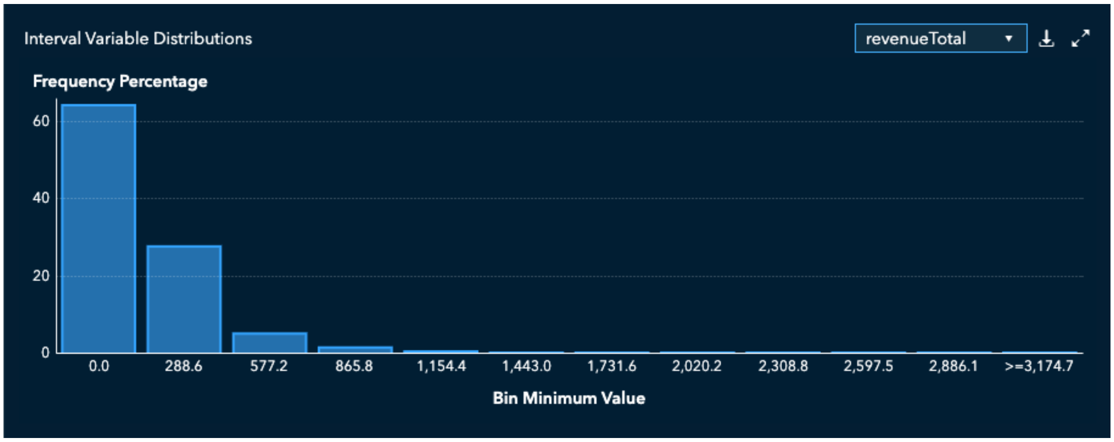
SAS photo 23: recchrge distribution



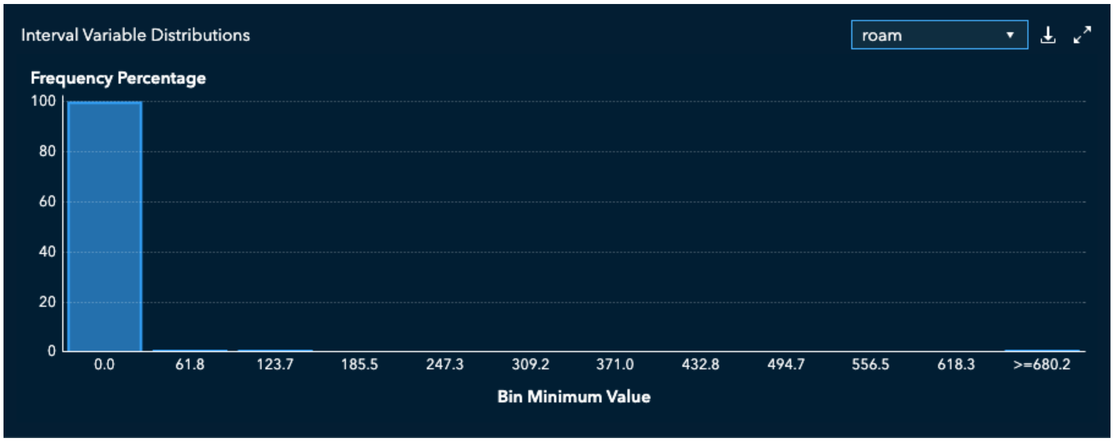
SAS photo 24: revenue distribution



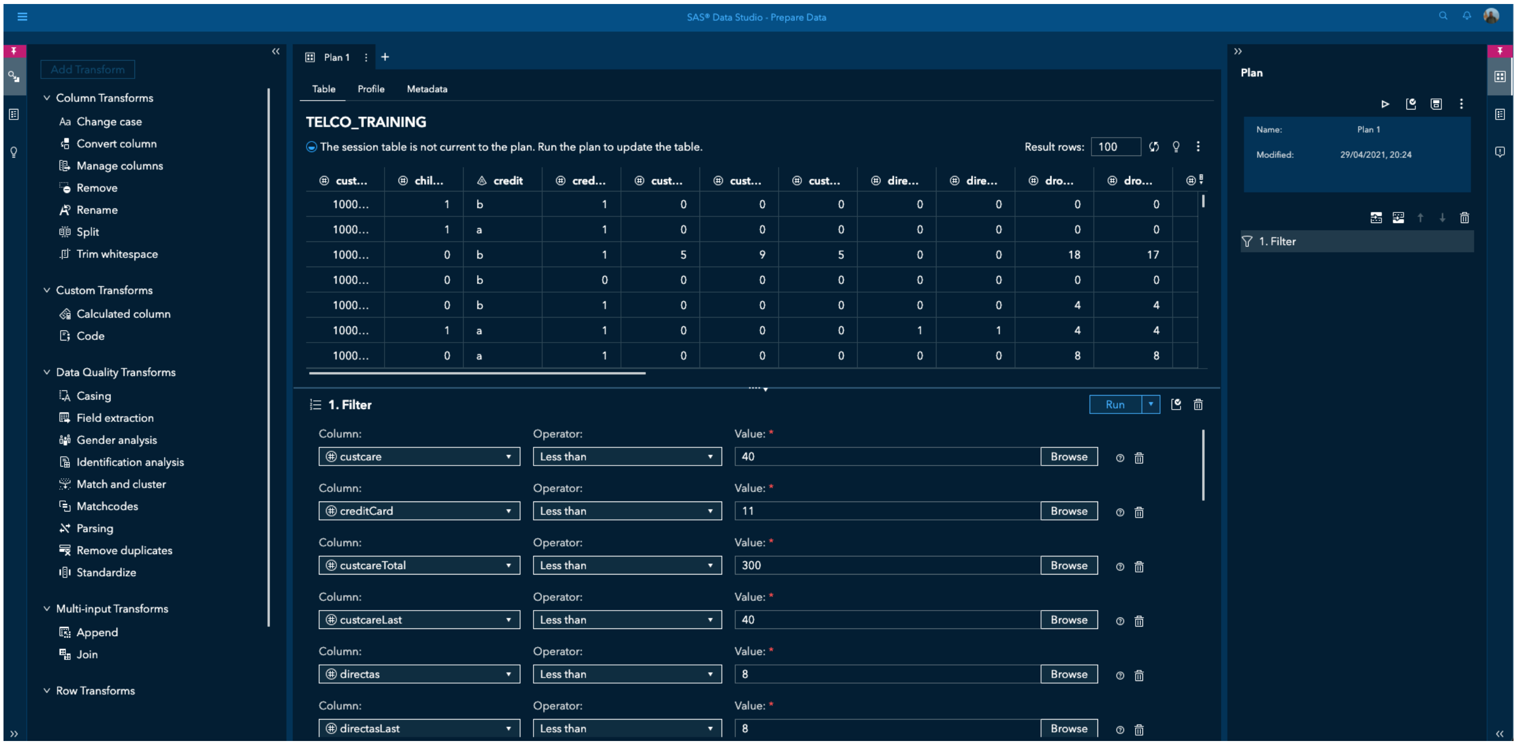
SAS photo 25: revenueChange distribution



SAS photo 26: revenueTotal distribution



SAS photo 27: roam distribution



TELCO_TRAINING

The session table is not current to the plan. Run the plan to update the table.

Result rows: 100

cust…	chil…	credit	cred…	cust…	cust…	cust…	dire…	dire…	dro…	dro…
1000…	1	b	1	0	0	0	0	0	0	0
1000…	1	a	1	0	0	0	0	0	0	0
1000…	0	b	1	5	9	5	0	0	18	17
1000…	0	b	0	0	0	0	0	0	0	0
1000…	0	b	1	0	0	0	0	0	4	4
1000…	1	a	1	0	0	0	1	1	4	4
1000…	0	a	1	0	0	0	0	0	8	8

1. Filter

Column:	Operator:	Value: *
custcare	Less than	40
creditCard	Less than	11
custcareTotal	Less than	300
custcareLast	Less than	40
directas	Less than	8
directasLast	Less than	8

Name: Plan 1

Modified: 29/04/2021, 20:24

SAS photo 28: Using the filtering in SAS to fix the data's outliers

Appendix D – Meeting Minutes & Declaration

Workload Distribution

Name	Responsible for:	Delivery to:	Signature
Vasileios Gounaris Bampaletsos	SAS code, data quality, technical parts	26 April	40314803
Gena Keenan	Introduction, discussion and recommendations, final edits	23 April	40208341
James Taylor	Exploration of results, discussions, report structure, finals edits	04 may	40312691
Linghe Tian	Clustering, general support	05 may	40320019
Zhiying Zhu	Marketing approach	25 April	40309491

Meeting Minutes

Meeting Date	Subject	Duration
12/04/2021	Meet with group for the first time. Discuss about the assignment. Split the work.	51 minutes.
19/04/2021	Discuss about structure. Future plans. Literature research.	20 minutes
26/04/2021	Discuss about data quality, marketing strategies, clusters	15 minutes
03/05/2021	Discuss about final model and further actions.	1 hour and 3 minutes
06/05/2021	Discuss about writing structure, prepare the writing report for uploading	21 minutes

Declaration

We confirm that the contribution to this assignment from each member was equal, the work is all our own and we agree to be awarded the same mark.

Signed:

James Taylor

Gena Keenan

Vasileios Gournaris Bampaletsos

Linghe Tian

Zhiying Zhu