TELCO ABC MARKETING STRATEGY REPORT

DATA MANAGEMENT AND MINING

SHIVANGI RAI (19200207)
PINKAL PRAJAPATI (19200050)
HANG NGUYEN (19202198)

Assessment Submission Form

Student Names	Shivangi Rai, Pinkal Prajapati, Hang Nguyen
Student Numbers	19200207, 19200050, 19202198
Assessment Title	Data Management and Mining Report
Module Title	Data Management and Mining
Module Co-Ordinator	Aoife D'Arcy
Tutor (if applicable)	
Date Submitted	24-04-2020
OFFICE USE ONLY **Date Received**	
OFFICE USE ONLY **Grade/Mark**	

A SIGNED COPY OF THIS FORM MUST ACCOMPANY ALL SUBMISSIONS FOR ASSESSMENT.

STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED.

Procedures for Submission and Late Submission
Ensure that you have checked the School's procedures for the submission of assessments.
Note: There are penalties for the late submission of assessments. For further information please see the University's ***Policy on Late Submission of Coursework***, (http://www.ucd.ie/registrar/)

Plagiarism: the unacknowledged inclusion of another person's writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism designed to impose sanctions that reflect the seriousness of University's commitment to academic integrity. Ensure that you have read the University's ***Briefing for Students on Academic Integrity and Plagiarism*** and the UCD ***Plagiarism Statement, Plagiarism Policy and Procedures***, (http://www.ucd.ie/registrar/)

Declaration of Authorship
I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

Names: Shivangi Rai, Pinkal Prajapati, Hang Nyugen

Date: 24-04-2020

EXECUTIVE SUMMARY

> Focus on the right message
> for the right people
> at the right time.

As a result of high customer churn rate that continues to rise, Telco ABC aims at improving its communications strategy in an attempt to reverse this trend.

This report is an attempt at building marketing strategies by identifying appropriate and effective communication methods suited to customer profiles. It explores data with two prime perspectives-

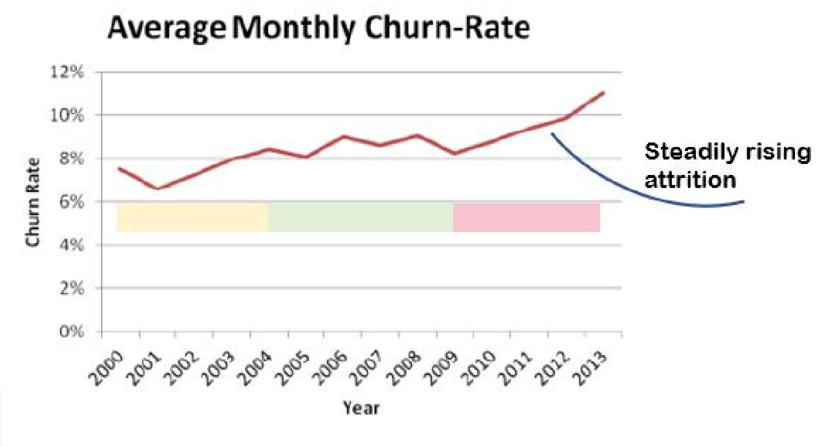
- Analyzing churn factors and building effective models to know who is at risk.
- Identifying customer segments to build the right strategies of customer engagement.
- Extracting the most valuable customer segments to determine how they can be kept happy so that such a profitable segment may expand.

Using the best suited analytical approach, it then throws light on churn prediction with recommendations to curb it and suggests effective marketing strategies for customer base retention and possibly, growth .

The key finding of this report is that people with higher income, consuming more services are Telco ABC's biggest cluster. This is the segment that can be expanded by regular communications via promotional messages, joining offers and appealing customer service options.

THE CONCERN IN A GIST

Depending on the market, **bagging new customers could be anywhere from 5 to 25 times more cost-incurring than holding on to an existing one**, reported a Harvard Business Review. Given the direct impact of that on revenue, wisdom would certainly lie in putting in that effort to retain old customers – which would still be modest – considering the rewarding revenue versus going about getting new customers to maintain the customer base. Telco ABC with a high attrition rate, is battling then not just declining revenue but also a fleeting customer base. The imminent impact of this – a poor public image and loss of trust.



DO WE SEE POTENTIAL?

Given the ubiquity of smartphones and the massive impact of social media on customer behavior, the ease with which customers can access various communication channels, it is a smart move to tap this potential to retain existing customer base and expand it. to An insight into customer profiles can help identify high churn rate areas as well as retention segments, deploy appropriate strategies and drive revenue.

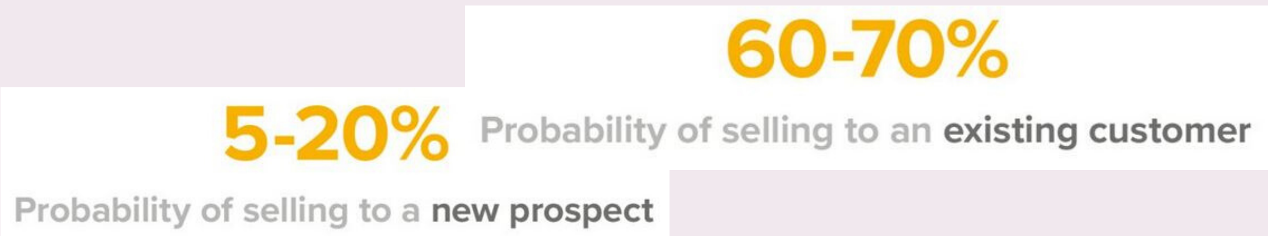
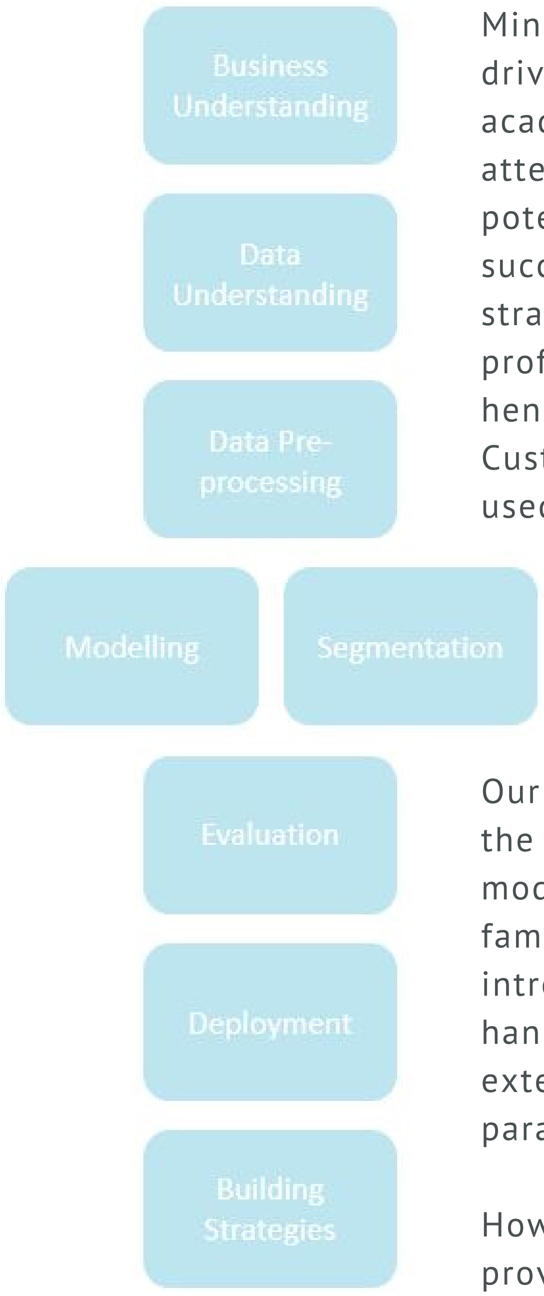


Image: Marketing Metrics claims on difference in probability of selling to new versus existing customers

INITIAL MARKET RESEARCH

Given the standardization in every business segment and the fact that organizations in each sector are connected by services they provide and the markets they target, we set out by understanding key Customer Relationship Management strategies and the role of Data Mining in maximizing growth by being data driven. This gave us both business and academic perspectives, in that it brought to attention successful strategy bases and the potential that data driven analysis had on successful outcomes. It stands out that all strategies revolve around tapping customer profile information to cater customized and hence effective, solutions. We identified the Customer Loyalty and Retention KPIs most used to build our research.



THE APPROACH

Our initial choice of software was Python given the extent to which descriptive analysis and modelling can be performed with it and our familiarity with it. That was until, we were introduced to SAS - a software adept in handling data, categorizing it and providing extensive insights for easily customized input parameters.

How data turned into information capable of providing insights is what we shall cover here after. The journey from the problem to solution building is displayed to the left.

DATA SPEAKS!

Given this is data driven strategy building, here is a look at the data exploration done.

DATA AT A GLANCE

Information around the 30 attributes was available for an approximate 10,000 customers. (around 8K, used for training models)
(See Appendix 1 for details on data)

These were divided into 7 categorical and 23 numerical attribute, the target variable being binary information on churn.

PRE-PROCESSING DATA

Knowing that an uncleaned, dirty data set can heavily affect analysis and skew model results, our primary task was to work on the initial data set.

Round 1 of data cleaning involved corrections to the data solely around value inconsistencies and missing data.

- Inconsistencies observed within “regionType” column – these were basically reporting inconsistencies around entries. They were all coded to align into three distinct categories ‘rural’, ‘suburban’ and ‘town’ to reflect correct and clear numbers.
-
- Missing values in “marry”, “occupation” and “regionType” columns, were aligned to 'unknown'

- Knowing that all analysis revolved around a high number of columns, we standardized column names for ease of identification.

- All columns with Binary values were converted to Booleans including the “Married” (yes/ no) values. (indicated also in above snapshot)

- To remove outliers we retained all data within a broad standard deviation range of 3.

- For all numerical values we defined meta data limits for columns with extreme values to filter out corresponding rows from analytics. The limits were set based on removal of 5% outliers if they fell far from the broader range.

- For an initial view of data, we added columns that grouped our numerical information into bands to give quick initial insights based on the quantiles

Temporary Column Added	Column Categories
Roaming Usage	nil, low, moderate, high
Call Drop Buckets	low, medium, high
Revenue Bucket	above average, below average

DATA CHARACTERIZATION

Financial status

This is a compound characteristic from variables "creditCard" and and "creditRating". Summary statistics indicate that over 60% of individuals have a credit card, and a majority majority of Telco ABC's customers have either an "aa" or "a" credit rating. Combinations split these into low credit rating (z, gy, de), moderate credit rating (c, b) or a high credit rating (a, aa) - including customers with a good credit score but no credit cards.

Familial Status

This characteristic is a potential parameter for deciding call plans. Data however was not complete against marital status and presence of children. A vast majority of customers however belonged to non-children households. Built from two binary variables "married" and "children" this helped us segment our customers.

Income level

With prospects of being a variable of high value, and well available data, the income spread of customers is denoted by a range from 0 to 9, with a median value of 5 (The mean income level of customers stands at around 4.28). Notably, the percentage of customers in income band 0 is considerably high (25.77%), compared to others.

These might either cover students, senior dependents and unemployed professionals, or are cases of absence of income reveal. From the marketing perspective, this attribute could be split into three categories namely High (6 – 9), Middle (3 – 5) and Low Income (0 – 2).

Roaming service

Knowing that revenue from roaming events is high, this data can help build attractive plans. A quick glance shows that most users avail of roaming services under 20 times with only a few showing significant roaming events against their usage.

Occupation

About 74% of the data available had no information on this. Although demographically a key aspect, considering the lack of information available, its influence to churn prediction was minimal.

Region

Approximately 60% Telco ABC’s customers live in suburban areas, with only about 9% spread in rural areas. While an important variable, the missing data percentage is being high, the region information in each customer segmentation is not much more than a guess.

Usage Minutes

The average minutes spent by users on call can help in spreading out plans to cater to segments more around the mean leading to higher subscriptions. With a mean of about 500 minutes and a significant variation of about about 550 on either sides.

Customer care calls, Directory Assisted Calls, Drop Calls (Last month)

Being inter-related and key parameters that throw insights on service quality, these were studied in dependency of one another.

Revenue Contribution (Last month & Total)

Given that customers value directly associates with the amount of revenue they generate, this column proved of utmost importance. It helps identify long-term customers from those who simply enroll in response to freebies and super-saver plans and likely churn once the benefit duration runs out. Most of the customers generate a revenue between 0 and 150, with significantly low number under 5 per customer - which translates well for the organization.

(Kindly see Appendix 2 for all descriptive statistic plots and parameter values, including the code for deriving it.)

ATTRIBUTE VALUATION

Our data is bound to contain a large spread of predictors. While some of these are useful, quite a few are not. SAS provides the "Variable Importance" feature that can give out the most useful predictors based on importance attached to variables - these will form the set that helps predict the response variable. This step is key in any machine learning approach and upon carrying it out for our data, below observations came to knowledge.

The 3 charts from top to bottom indicate the relative worth of variables in decreasing order.
Knowing that revenue would have a high importance, and since the first chart showed much lower worth of other variables in comparison, the next two charts simply zoom in on the importance of the other factors by blinding out those already considered.
A parallel confirmation of attribute worth was also done in python to confirm insights and obtain the list of key variables impacting the target.

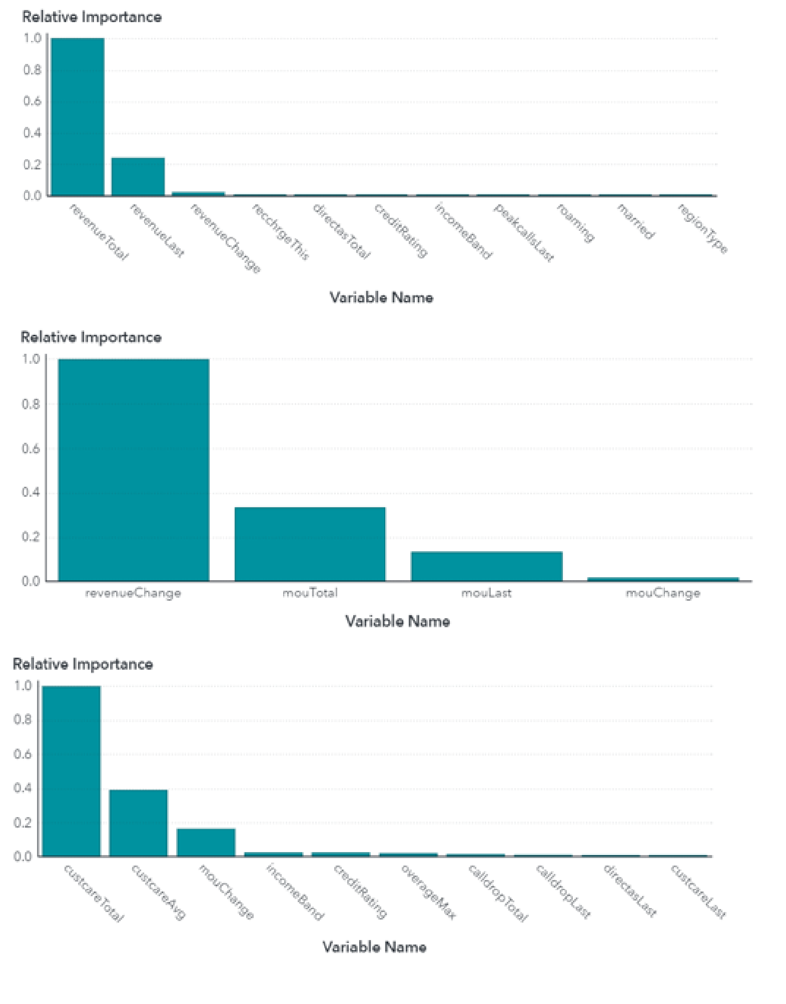


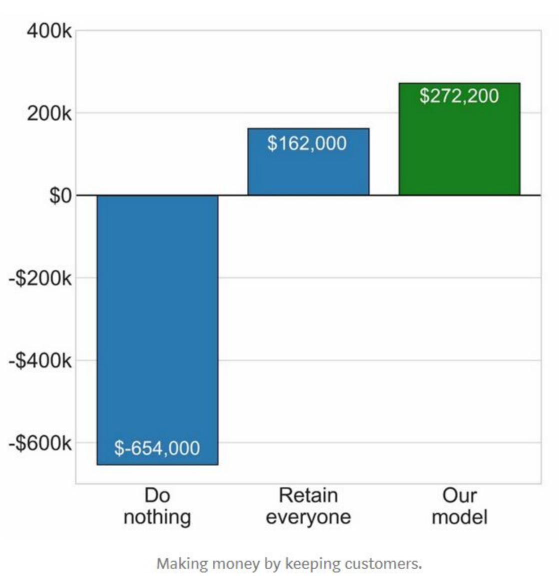
Image: A snapshot from our SAS data exploration results for each instance

EXPLORING ANALYTICAL APPROACHES

The approach taken determines the direction of analysis and the validity of the outcome. Operators incorporate advanced machine learning algorithms to mines of data to discover insights. Variables that appear otherwise dormant, end up in valuable buckets - one such method is feature discovery. Apart from this, the relativity among attributes and their collective importance also comes to light. What these point to are buckets of customers with distinct behaviors.

This points to the concept of segmentation and micro-segmentation. It has been established how blind application of retention techniques to all customers yields lower revenue as opposed to tailored-targeting of customers.

The picture to the right shows results from an experiment of applying predictive modelling techniques to curb attrition rate. It shows how a good model can save revenue as opposed to blindly utilizing retention costs or doing nothing at all.



Analytical retention techniques v/s other options.
Source: towarddatascience

While in Decision Trees, a tree like structure represents related choices based on decision rules (Berry & Linoff, 2004), in Gradient Boosting, an incremental technique converts weak trees into strong ones while Logistic Regression chooses a least restrictive but best-fitting model to describe relationships between several inputs and a binary target.

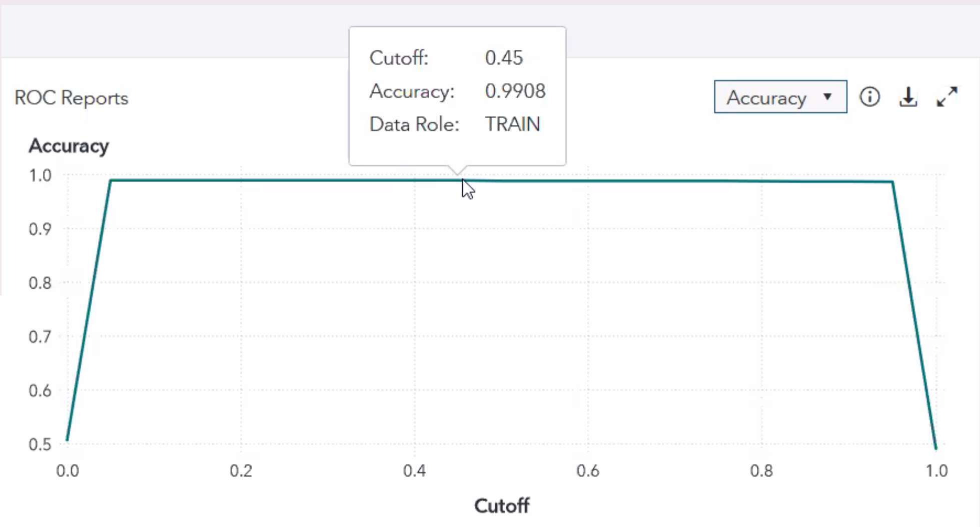
We explored each of these models to identify the one best suited to our case.

PREDICTIVE MODELLING

Prior to deployment, the data was filtered and given to different predictive models like Decision Tree, Gradient Boosting which builds a sequential series of decision trees and Logistic Regression for a binary target, in our case “churn”. All of these models essentially split the data into training, testing and validation inputs.
The output of each of these models was then compared using the following parameters-

- Observing the Receiver Operating Characteristic curve which gives a representation of the trade-off between specificity and sensitivity.
- The Accuracy parameter which is a quick and valuable indication of model performance
- The model scaling the highest Area Under Curve
- Cumulative Lift Gain that measures the effectiveness of a predictive model.

Case studies under Towards data Science and Journal of Big data showed Logistic Regression to perform significantly well, and that reflected well on our data too, outperforming Decision Tree and Gradient Boosting.
The misclassification rate for Logistic Regression (0.003 for training and 0.006 for test) was significantly lower as compared to other model like Decision Tree (0.06, 0.05). The cumulative lift scores for the logistic regression model were 2.0159 (train) and 2.0163 (test), respectively.
This model with the lowest misclassification rate and highest accuracy proved to be our champion model and was deployed on our test data.



Accuracy rate for the Logistic Regression model on Training Data

Model Comparison

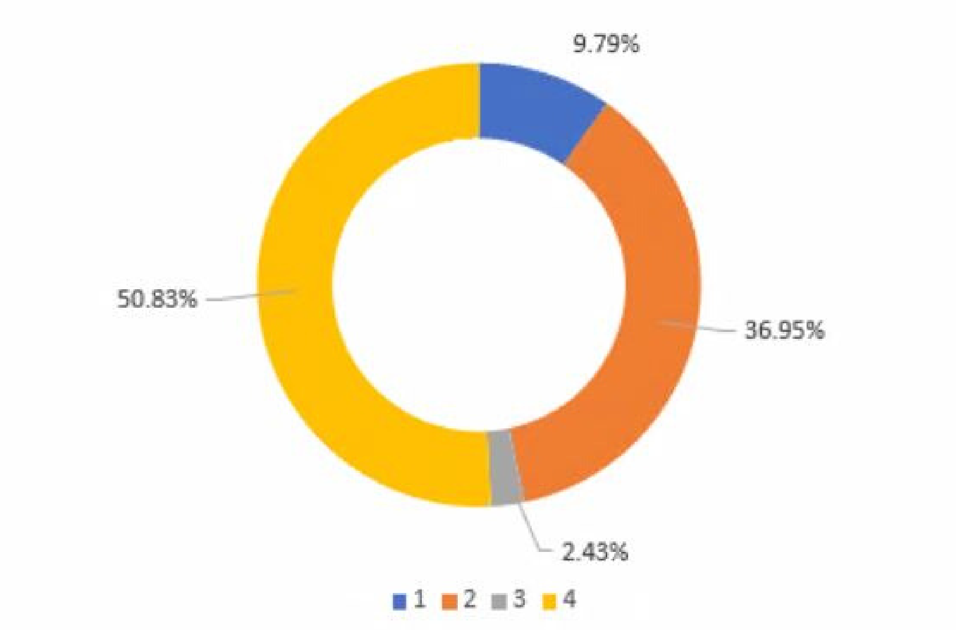
Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
	Logistic Regression	Logistic Regression	0.9950	0.0037
	Gradient Boosting	Gradient Boosting	0.9775	0.0125
	DT : Gini	Decision Tree	0.9474	0.0312
	DT : Info Gain	Decision Tree	0.8896	0.0587

Model Comparision Results

SEGMENTATION

> The customer expects you to have knowledge of their stuff, not just your stuff.
> -Jeffrey Gitomer

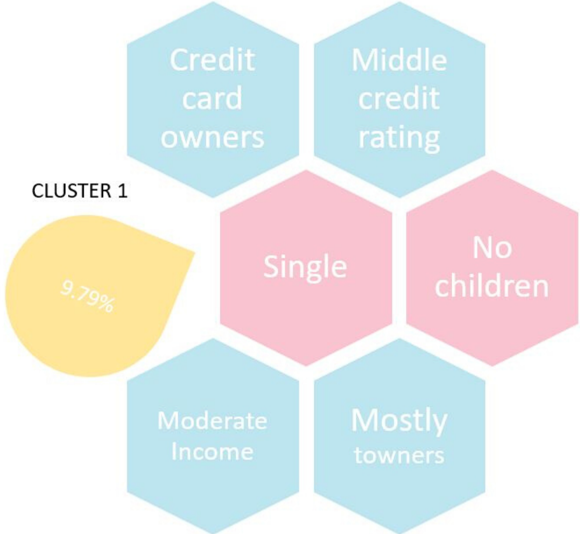
In order to deploy the right strategy to curb churn rate, it thus becomes key, to map the dissatisfaction among customers to their spending, loyalty and and profiles. Exploring causes for a dissatisfied customer for instance would bring us to call drops and frequent reaching out to customer care. These would likely indicate a negative steer in revenue.
By selecting attributes that define a customer profile (revenue, credit, credit card and familial status) the below key clusters showed up.



A snapshot of the clusters created

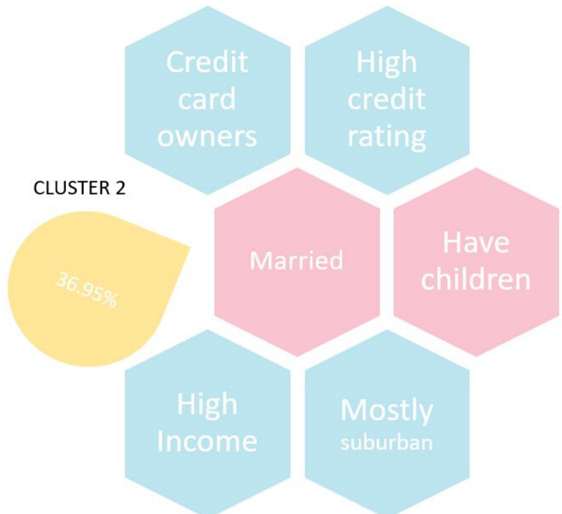
The Spice (9.79%)

Let not the size of this cluster fool one - although a modest 10%, it is the most profitable segment for Telco ABC, considering that revenue and the recurring bundle charge from this group is always highest among four customer groups, 68.4 and 49.4 respectively for the last month.



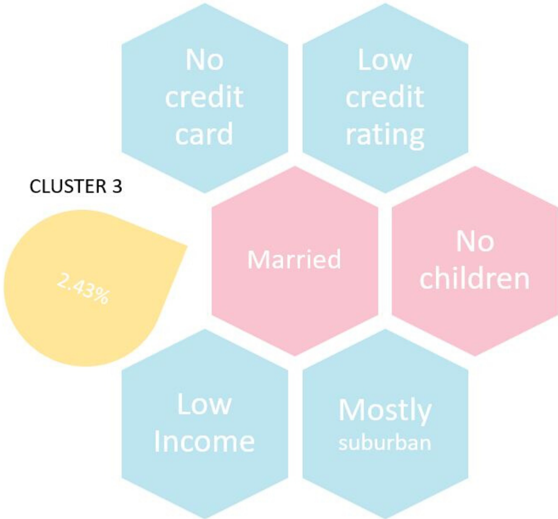
The Decorative Lemon Grass (36.95%)

Sure this segment is noticeable and has a fancy share , but it adds nothing to satiate our appetite - i.e. the revenue. The least profitable with lowest recharges to recurring bundles, 54.97 and 43.09 respectively for the last month. The phone call duration as well are half that of cluster 1.Although, notably calls in last 6 month duration are significantly higher than each of the other groups.



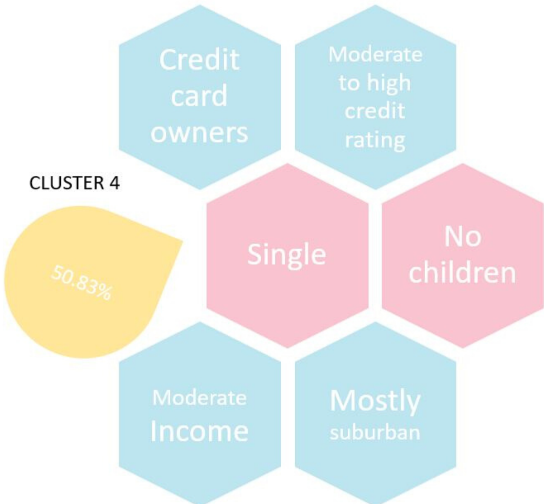
The Salt (2.43%)

Although having lower finance status, they show a positive revenue change (0.14%), highest among all segments. Revenue from this group also ranked second for the last month and third for the last six months. Ranking 2 in recurring bundle charge and 1 in peak calls, higher than average minutes, they certainly make a difference for their amount - like the salt. This segment has the least roamers and might well be office goers.



The Gravy (50.83%)

Highest in proportion and with the most roaming events (20% over segment 1) - this segment moves. Average phone calls and revenue, and a low peak duration In the last six month, their peak call usage is low - which explains them being roamers.



THE PROPOSED STRATEGY

> In marketing I've seen only one strategy that can't miss - to market to your best customers first." - John Romero

The Spice being our most valuable segment, providing them quality service will only make them happier. Occasionally sending them rewarding offers on getting their acquaintances to enroll, can work well since these are bound to have a good word to share about us.

T**he Salt** is our segment with above average performance and show potential (increasing revenue change) despite the low incomes. Given their highest bundle recharge trend, they could be called to learn of better bundle expectations. A bundle reflecting their average talk-time would be lucrative. It would be noble to create and offer affordable plans to make empower this tiny segment which probably has students and dependents.

The Gravy being the largest segment likely represents the middle and upper middle class customers. Feedback calls and social media polls to samples in this section can provide insights on what could be made better. If successfully understood, a positive change in revenue from this section can move the market share and profits largely.

The Lemon grass notably, has individuals with high revenue. Given that their purchasing power is high, these are likely deterred by poor service and would rather opt for better quality high-end services. A check on quality coupled with increased customer service and better data plans can churn this crowd's potential.

APPENDIX

APPENDIX 1: DATA

Variable	UpdatedVariable	Description
customerID	customerID	Customer ID
children	children	Indicates if children are present in the household {true, false}
credit	creditRating	The customers credit rating {a, aa, b, c, de, gy, z}
creditCard	creditCard	The customer owns a credit card {true, false}
custcare	custcareAvg	Average number of calls to customer calls in the last 6 months
custcareTotal	custcareTotal	Total calls to customer calls in the last 6 months
custcareLast	custcareLast	calls to customer calls in the last month
directas	directasTotal	The number of directory assisted calls made in the last 6 months
directasLast	directasLast	The number of directory assisted calls made last month
dropvce	calldropTotal	The number of calls dopped in the last 6 months
dropvceLast	calldropLast	The number of calls dopped the last month
income	incomeBand	The cutomer's income {0 - 9}
marry	married	The customer's marital status {yes, no, unknown}
mou	mouLast	Number if minutes last month
mouTotal	mouTotal	The total number of minutes used in the last 6 months
mouChange	mouChange	% change in minutes
occupation	occupation	The occupation of the cutomer
outcalls	outcalls	The number of calls made
overage	overageThis	The number of minutes over the customer's bundle used this month
overageMax	overageMax	Max overage
overageMin	overageMin	Min overage
peakOffPeak	peakcallsTotal	The total number of peak calls made the last 6 months
peakOffPeakLast	peakcallsLast	The total number of peak calls made last month
recchrge	recchrgeThis	The recurring bundle charge this month
regionType	regionType	The type of region in which the customer lives {rural, suburban, town}
revenue	revenueLast	Reveue from customer last month
revenueTotal	revenueTotal	total revenue in the last 6 months
revenueChange	revenueChange	% change in revenue
roam	roaming	The number of roaming events in the time period
churn	churn	Flag indicating if the customer has churned

Figure 1: Data Dictionary with Updated Column Values

APPENDIX 2: DESCRIPTIVE STATISTICS ON DATA

In [51]:
```python
df.describe()
(df.describe()).to_excel("quantileData.xlsx")
```

Out[51]:

	customerID	custcare	custcareTotal	custcareLast	directas	directasLast	dropvce	dropvceLast	income	mou	...	
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	...	10000
mean	1.049974e+06	1.746400	8.090200	1.727800	0.90930	0.882000	5.985100	5.924600	4.293600	521.887595	...	4
std	2.879841e+04	5.861068	31.586415	5.885046	2.28989	2.252284	8.654488	8.641677	3.139902	542.907394	...	10
min	1.000001e+06	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	...	(
25%	1.025200e+06	0.000000	0.000000	0.000000	0.00000	0.000000	1.000000	1.000000	0.000000	150.999321	...	(
50%	1.049833e+06	0.000000	0.000000	0.000000	0.00000	0.000000	3.000000	3.000000	5.000000	361.336280	...	(
75%	1.074990e+06	1.000000	6.000000	1.000000	1.00000	1.000000	8.000000	7.000000	7.000000	708.051098	...	42
max	1.099988e+06	376.000000	2253.000000	387.000000	55.00000	52.000000	114.000000	110.000000	9.000000	6494.324859	...	431

8 rows × 23 columns

In [52]:
```python
#Replacing spaces with null values in total charges column
df = df.drop(['occupation', 'regionType', 'customerID'], 1)
# drop last irrelevant column
df = df.iloc[:, :-1]
print (df.isnull().sum())
```

```
children        0
credit          0
creditCard      0
custcare        0
```

Figure 2: Descriptive Statitics using Python

Variable Summary

feature	count	mean	std	min	25%	50%	75%	max
children	1464	0.51	0.5	0	0	1	1	1
creditCard	1464	0.954	0.209	0	1	1	1	1
custcare	1464	0.956	2.831	0	0	0	1	52
custcareTotal	1464	4.32	13.572	0	0	0	1	206
custcareLast	1464	0.941	2.766	0	0	0	1	49
directas	1464	0.654	1.594	0	0	0	1	27
directasLast	1464	0.624	1.544	0	0	0	1	23
dropvce	1464	4.375	6.557	0	1	2	5	78
dropvceLast	1464	4.341	6.642	0	1	2	5	73
income	1464	6.275	2.037	0	5	6	8	9
marry	1464	0.779	0.415	0	1	1	1	1
mou	1464	393.726	452.773	0	103.32	262.394	527.941	6494.325

Figure 3: Descriptive Statistics on high worth variables

APPENDIX 2: PROCESSED DATA VISUALS

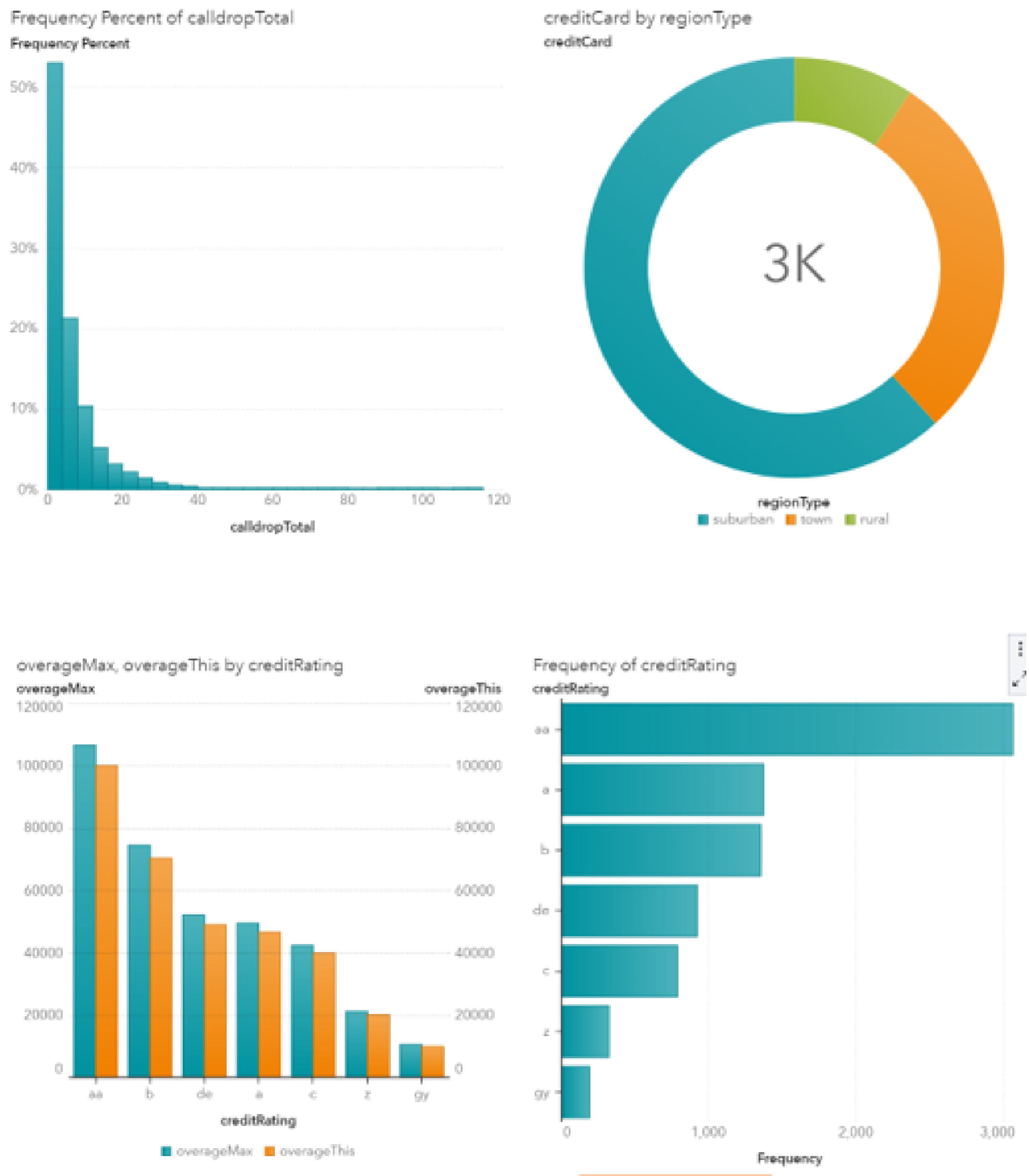


Figure 4: Data Insights Viduals (Set 1)

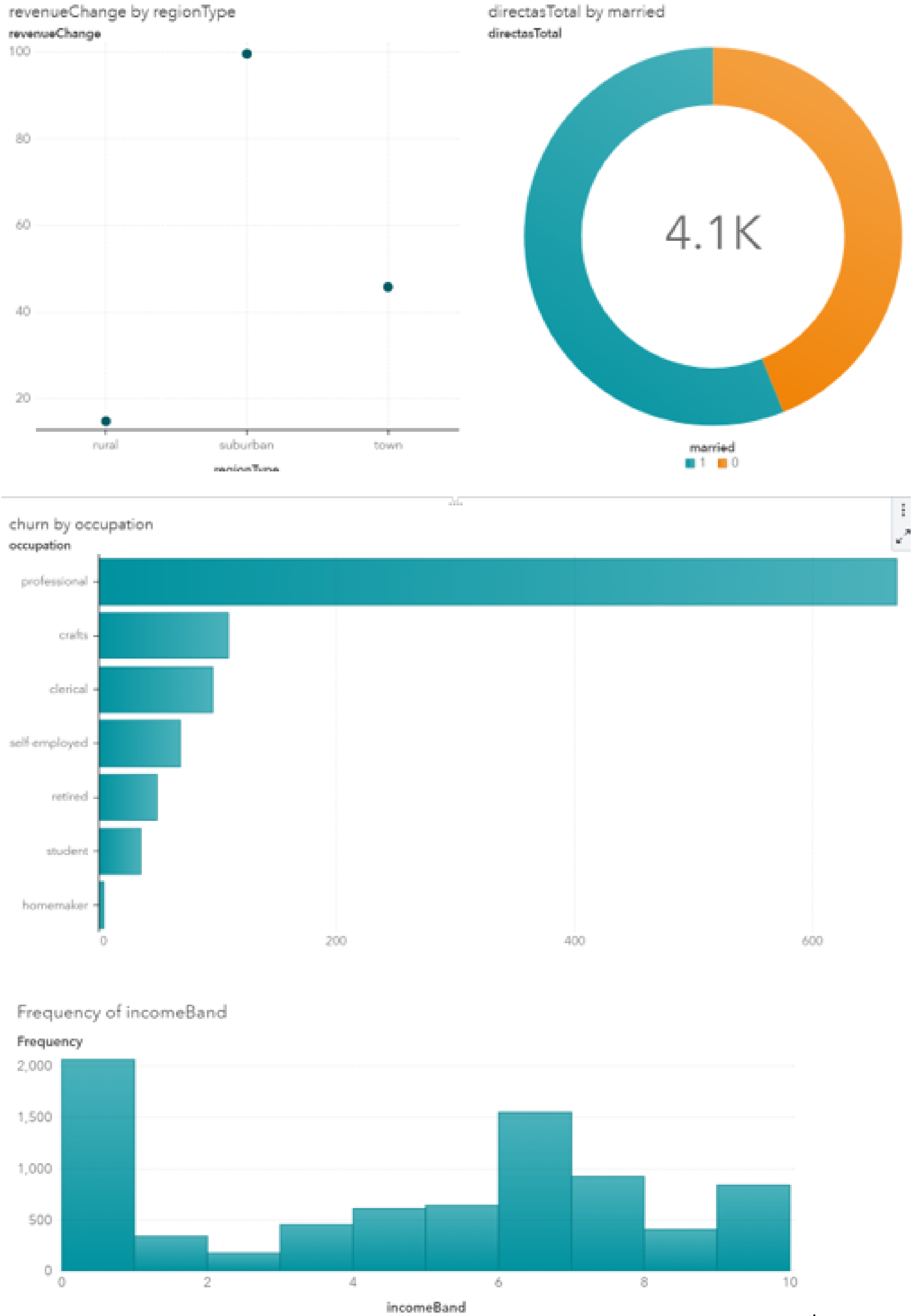


Figure 5: Data Insights Viduals (Set 2)

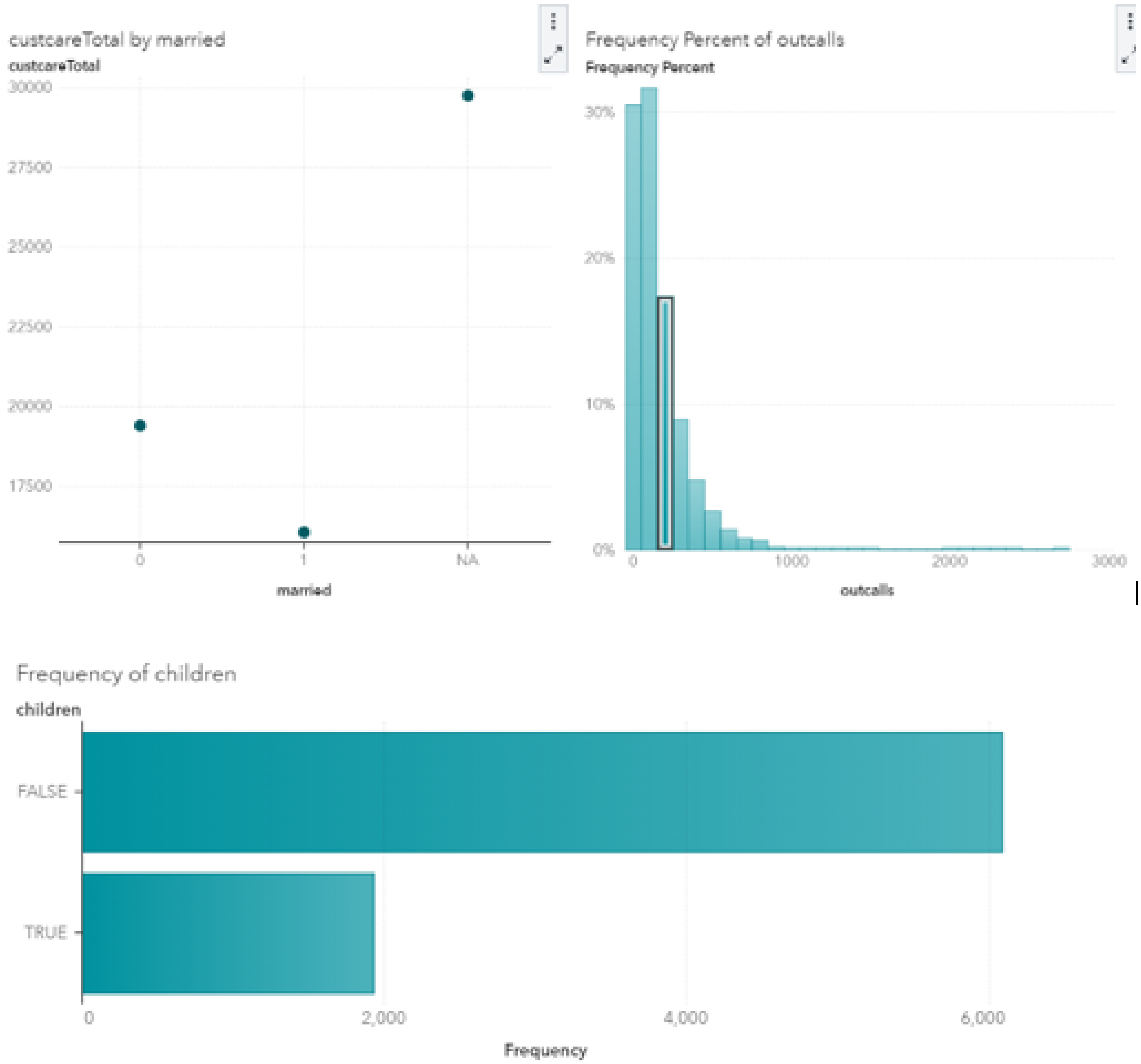


Figure 6: Data Insights Viduals (Set 3)

APPENDIX 3: PREDICTIVE MODELS

Target ...	Data Role	Partitio...	Formatt...	Sum of ...	Averag...	Divisor ...	Root Av...	Misclas...	Multi-Cl...	KS (You...	Area Un...	Gini Co...
churn	TEST	2	2	801	0.0030	801	0.0547	0.0037	0.0549	0.9950	0.9975	0.9950
churn	TRAIN	1	1	4,804	0.0003	4,804	0.0186	0.0006	0.0010	0.9988	1.0000	1.0000
churn	VALIDATE	0	0	2,401	0.0078	2,401	0.0886	0.0087	0.1352	0.9850	0.9933	0.9866

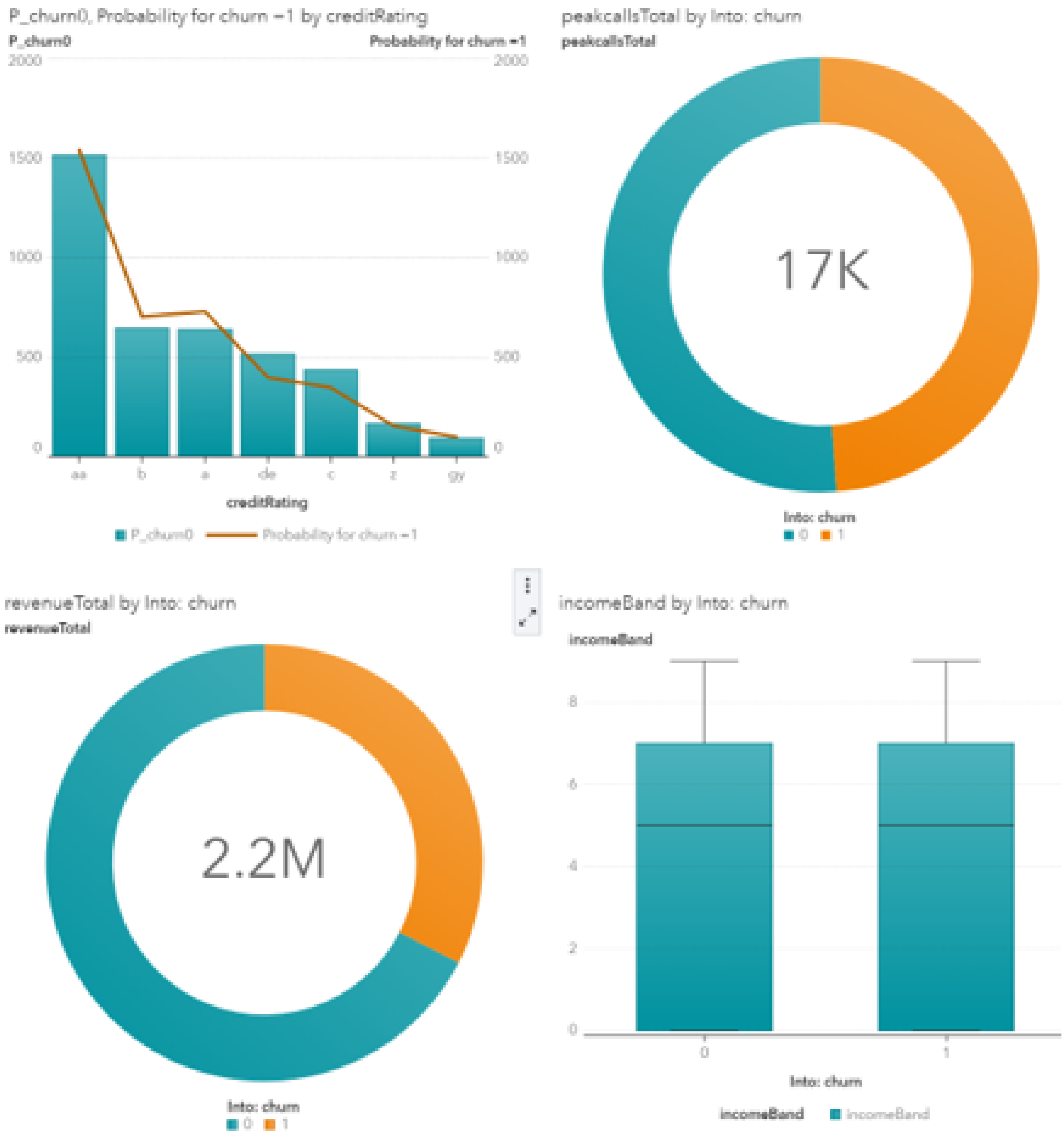


Figure 7: Logistic Regression Visuals

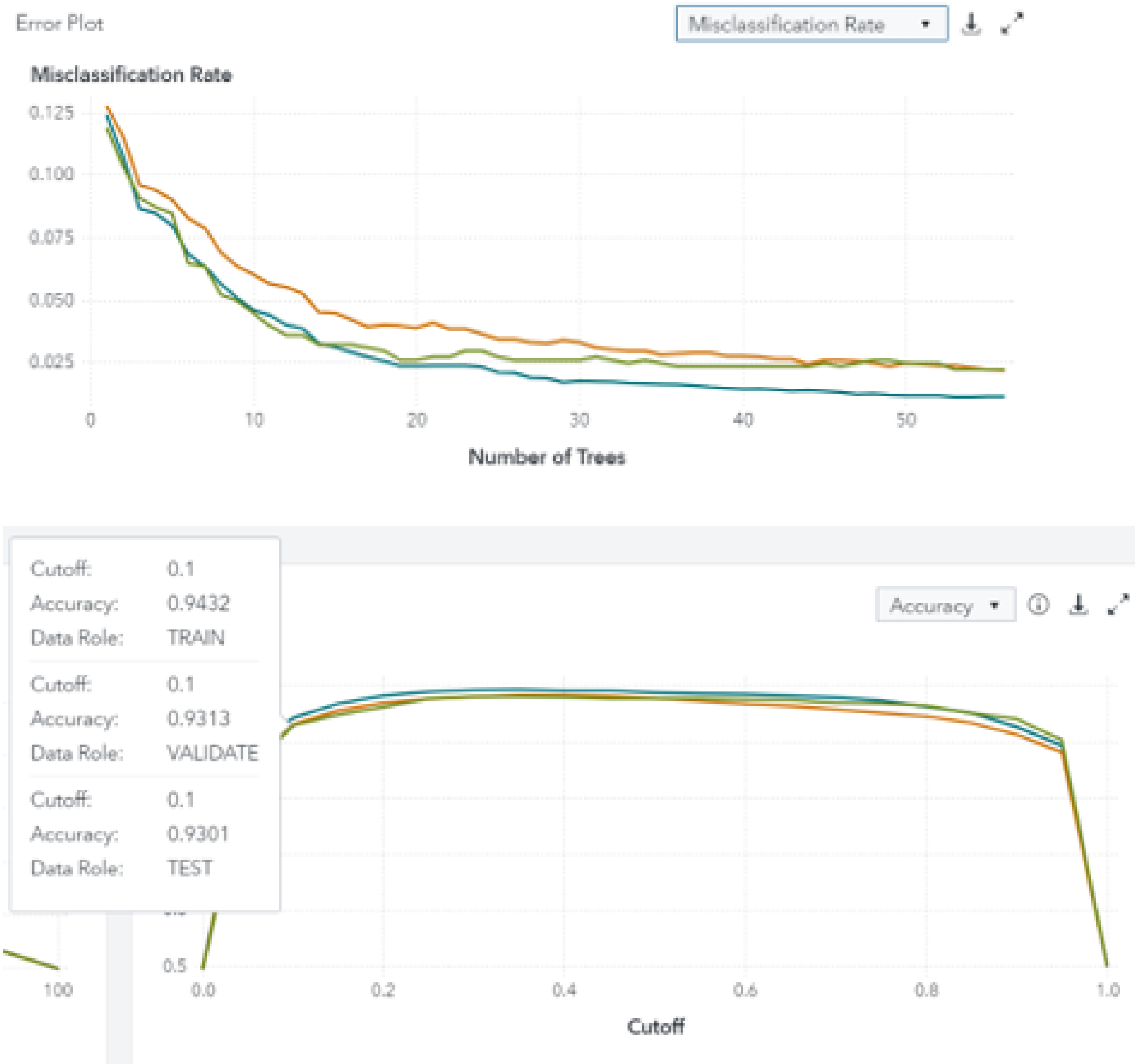


Figure 8: Gradient Boosting Visuals

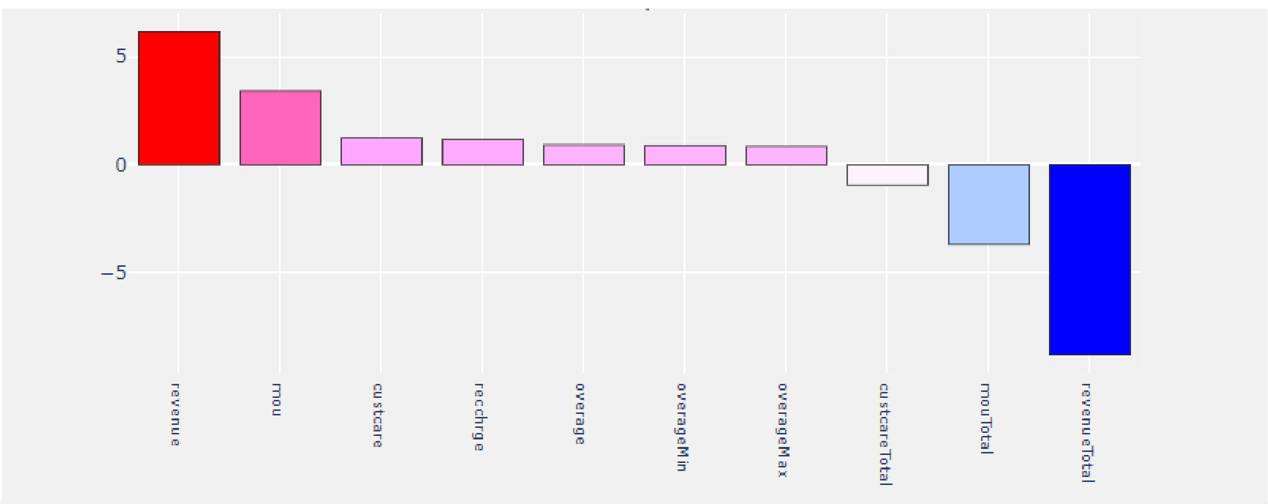


Figure: Feature Correlation to Target

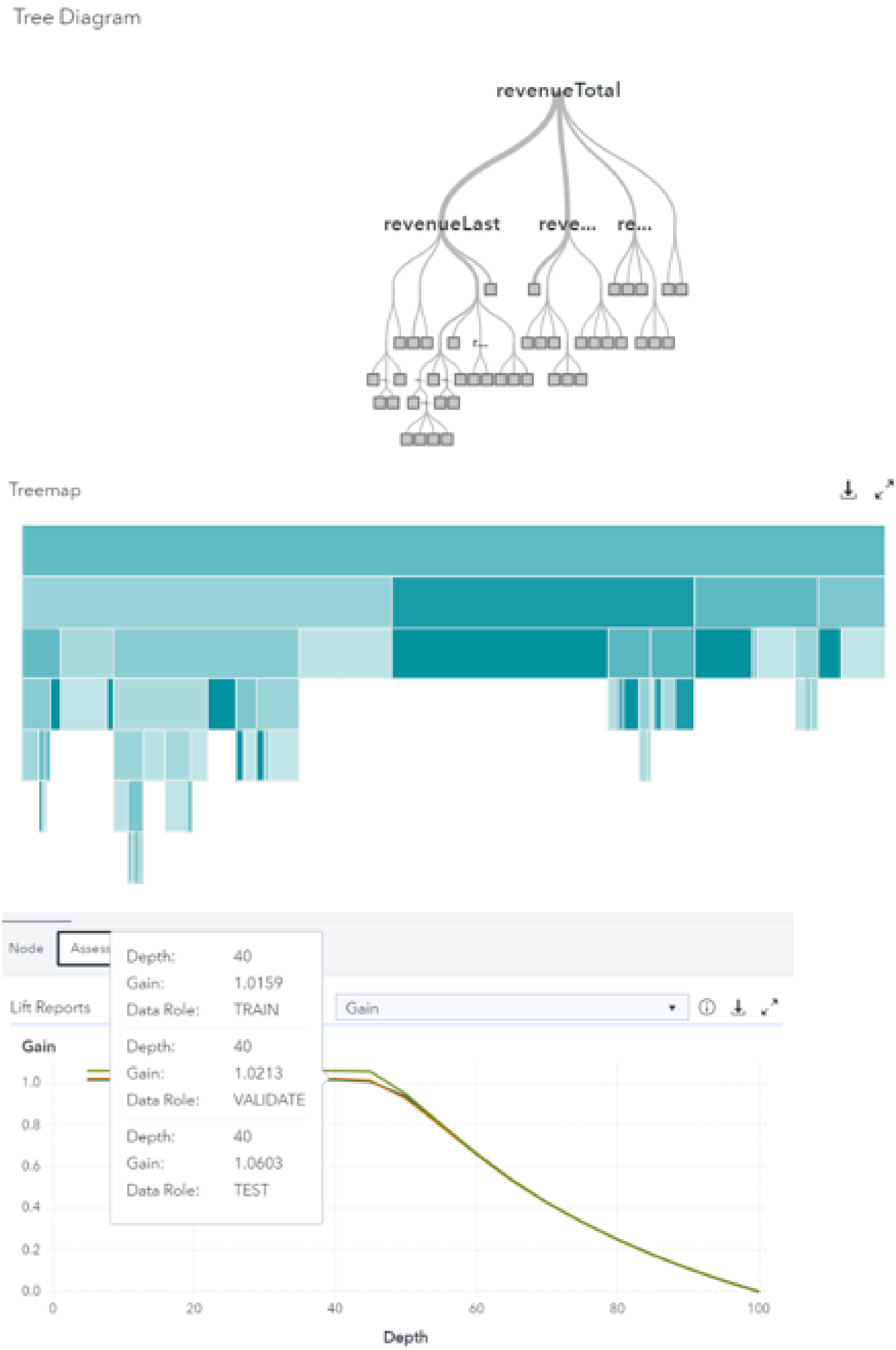


Figure 8: Decision Tree (Gini) Visuals

APPENDIX 4: MODEL COMPARE & DEPLOY

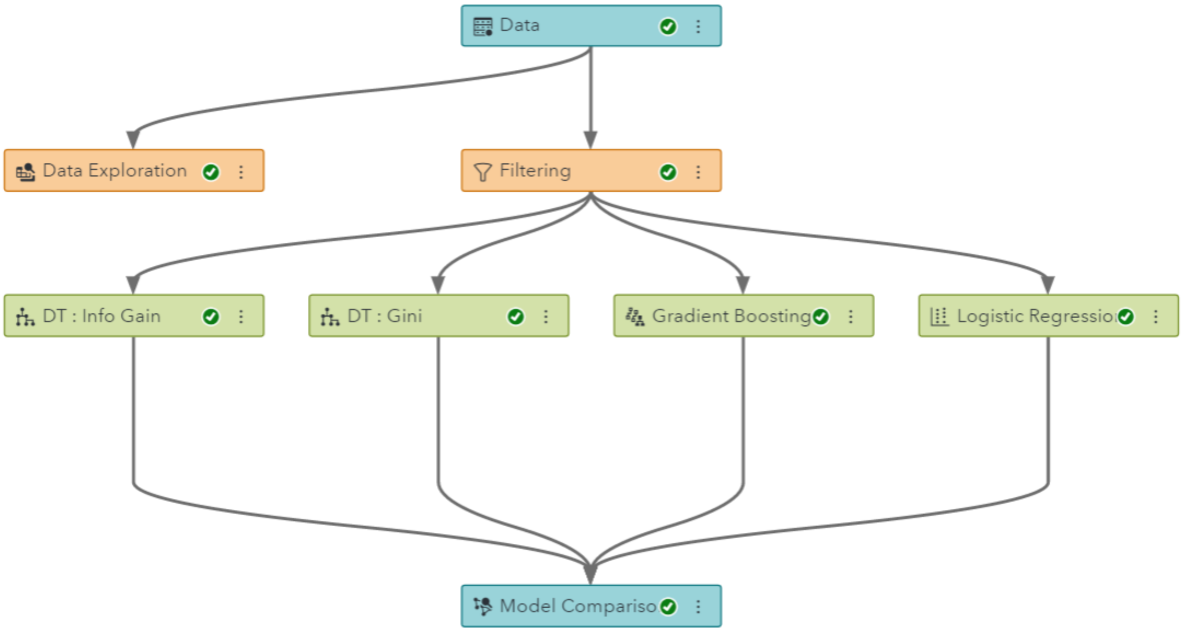


Figure 9: Pipeline

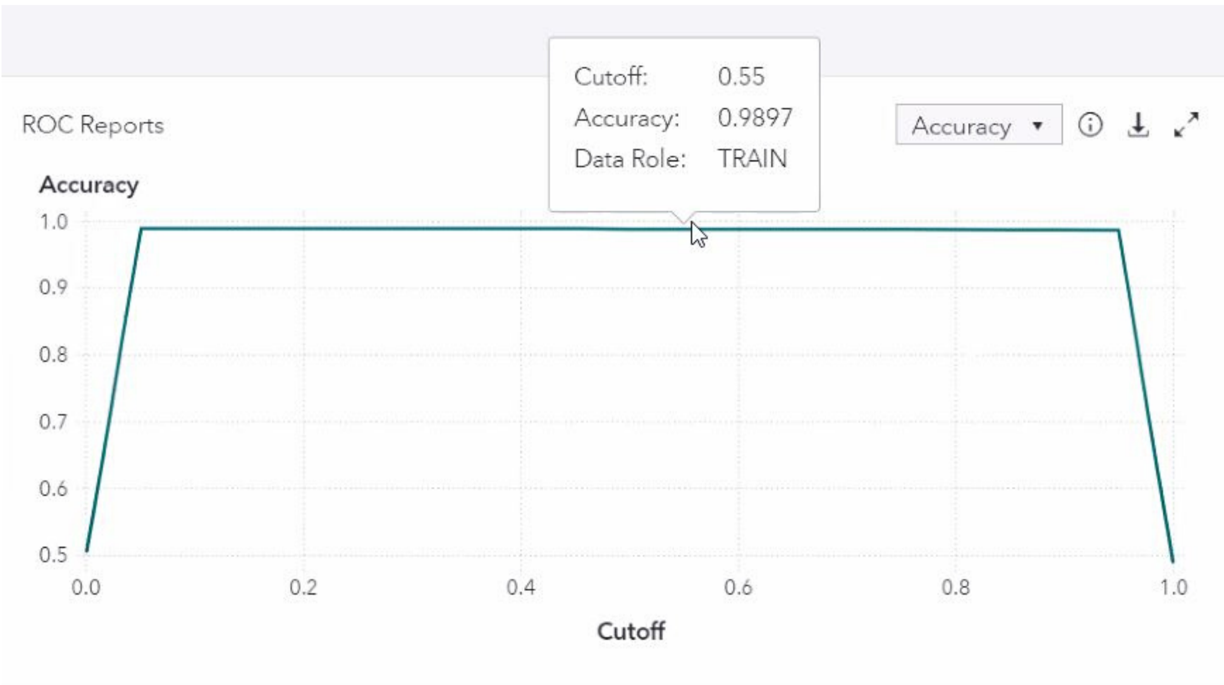
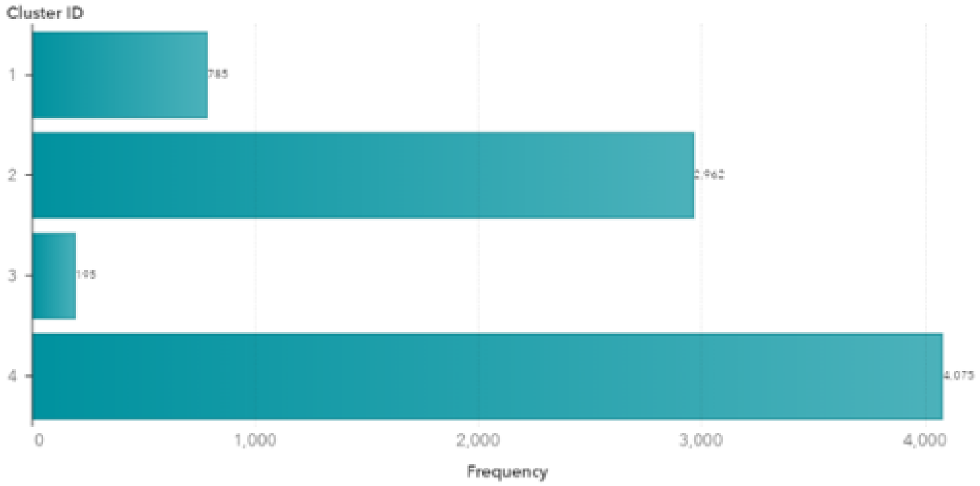


Figure 10: Accuracy postLR model deployment on Scoring data

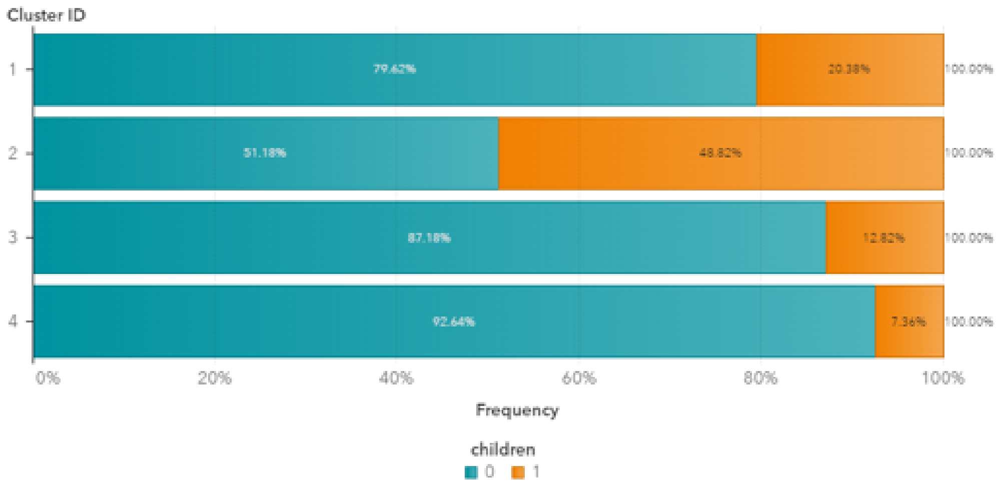
APPENDIX 5: SEGMENTATION

Variable Name	Root Variable N...	Role ↓	Measurement L...
churn	churn	TARGET	BINARY
occupation	occupation	REJECTED	NOMINAL
children	children	REJECTED	BINARY
PartInd	_PartInd_	PARTITION	NOMINAL
dmIndex	_dmIndex_	KEY	NOMINAL
custcareAvg	custcareAvg	INPUT	INTERVAL
directasTotal	directasTotal	INPUT	INTERVAL
creditRating	creditRating	INPUT	NOMINAL
directasLast	directasLast	INPUT	INTERVAL
creditCard	creditCard	INPUT	BINARY
peakcallsLast	peakcallsLast	INPUT	INTERVAL
incomeBand	incomeBand	INPUT	NOMINAL
married	married	INPUT	NOMINAL
mouChange	mouChange	INPUT	INTERVAL

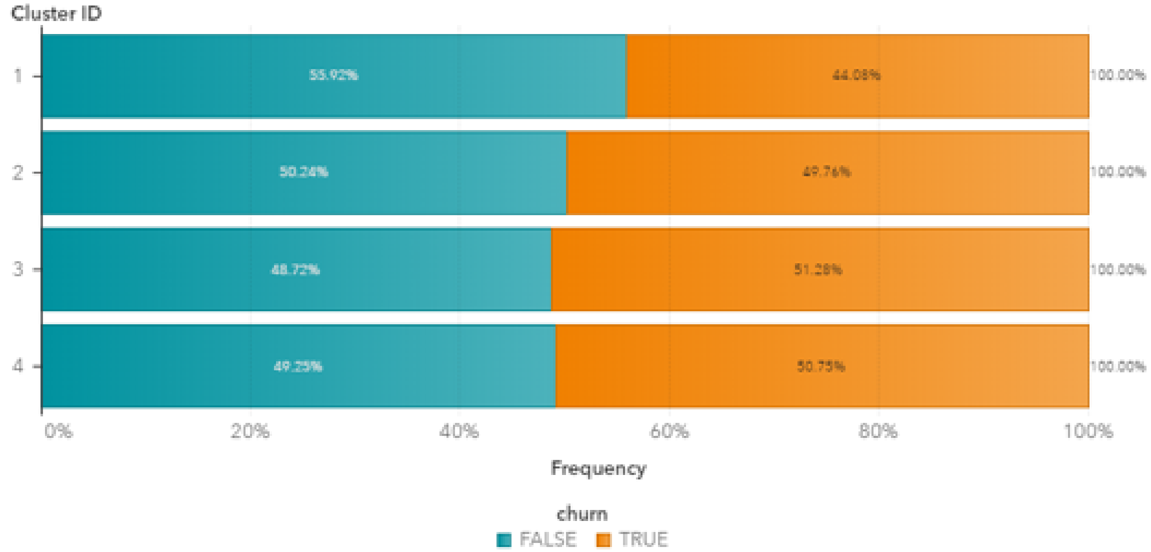
Frequency of Cluster ID



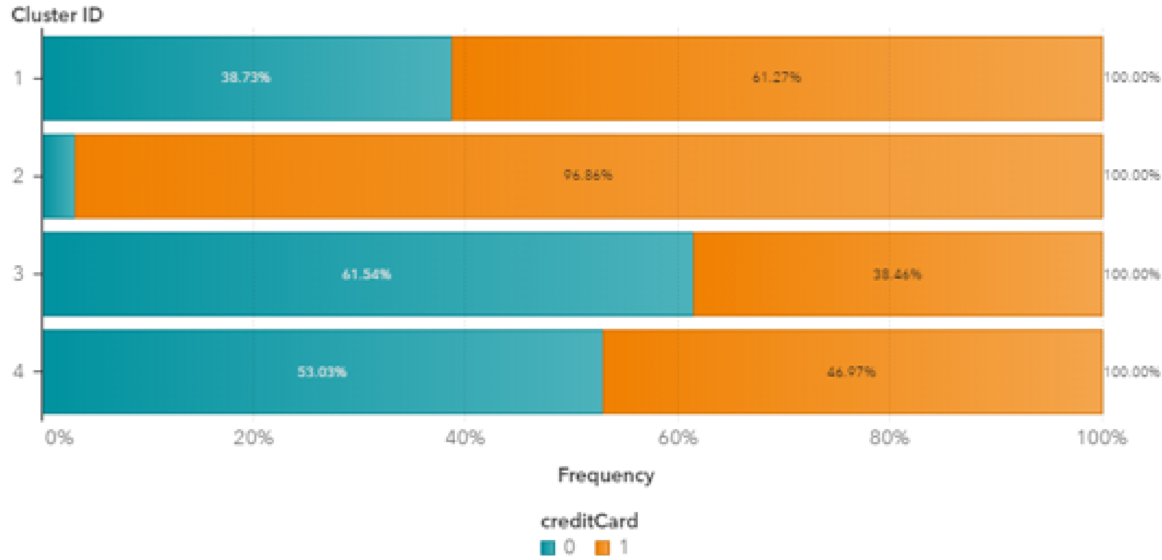
Frequency of Cluster ID grouped by children



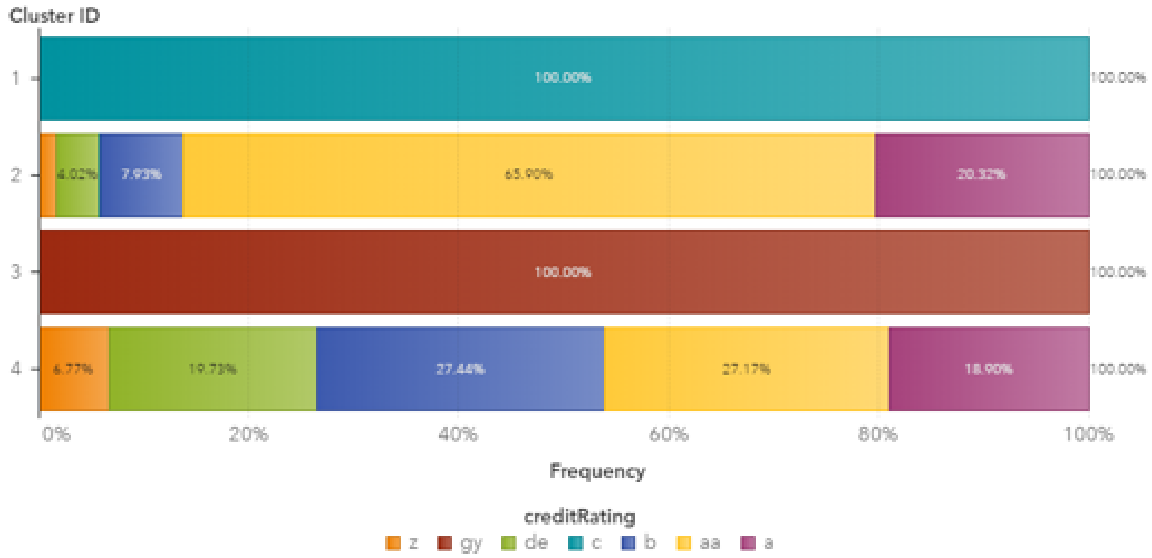
Frequency of Cluster ID grouped by churn



Frequency of Cluster ID grouped by creditCard



Frequency of Cluster ID grouped by creditRating



WORK LOG

Meeting 1 (March 6):

- Identification of different analysis approaches.
- Discussion made on various modelling approaches.
- Initial work on descriptive statistics – Python (Siu), R language (Pinkal), SAS + Excel (Shivangi)

Meeting 2 (March 24):

- Performed high level descriptive analysis using the different approaches.
- Developed Pseudo code in python for assessment.
- Created a draft of the data quality report.

Meeting 3 (March 06):

- Performed Data cleaning using Excel and Python.
- Identification of important variables and filtering them.
- Learned SAS environment and shared understanding with each other.
- Data preparation in SAS – ETL process.

Meeting 4 (April 12):

- Data exploration and Data pre-processing in SAS.
- Implemented different predictive models using SAS model building tool.
- Learned about model comparison and their performance indicator to decide the best predictive model. Implemented Scoring of data set
- using built in predictive models.

Meeting 5 (April 21):

Clustering and segmentation.
Discussion of various marketing
strategies based on different segments.

Meeting 5 (April 21):

- Report data points (Hang), Report framing and formatting (Shivangi), gathering supporting data and snippets (Pinkal)

Meeting 7 (April 24):

- Final review

REFERENCES

- Baier D., 2018, Important KPIs for the Telecom Industry, TGDaily, viewed 24 April 2020,<https://www.tgdaily.com/technology/importan-kpis-for-the-telecom-industry/>.

- Berry, M. & Linoff, G., 2014. ‘Data Mining Techniques for Marketing, Sales and Customer Relationship Management (2nd edition)’, Wiley, Indianapolis.

- Dahiya K. and Bhatia S., 2015, "Customer churn analysis in telecom industry", 4th International Conference on Reliability, Infocom Technologies and Optimization ICRITO, Trends and Future Directions, Noida, pp. 1-6.

- Deshpande, B. & Kotu, V., 2014. ‘Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner’, Elsevier, London.

- Gallo A., 2014, The Value of Keeping the Right Customers, viewed 22 April 2020, <https://hbr.org/2014/10/the-value-of-keeping-the-right-customers>.

- SAS Institute Inc., 2018, 8.2: Advanced Topics. Preprocessing by Unsupervised Learning, SAS® Visual Data Mining and Machine Learning, viewed 23 April 2020, <https://documentation.sas.com/?docsetId=vdmmladvugdocsetTarget=n1e4szcnv1f0fn1vsxhbzgdp1bb.htm&docsetVersion=8.2&locale=en>.