

**MUKESH PATEL SCHOOL OF
TECHNOLOGY MANAGEMENT
& ENGINEERING** TM

PROJECT REPORT

Subject: Predictive Modeling (PM)

Team Name: Data Wizards (Kaggle)

Stream: MBA Tech.

Submitted by:

Rohit Bhatia I006
Heenal Bhavsar I009
Ayush Madhani I031
Ishaan Mulki I034
Dhrumil Vadodaria I061

FACULTY: Dr. Shailaja Rego

1. Introduction

A telecommunications company can increase its revenue and engagement through data services like tailor-made mobile marketing campaigns. Through examining customer profiles comprising of age, occupation and education, we prioritize the targeting and are successful in this. Through applying predictive analytics, our main thrust is to reduce costs on acquisition and personalize the offers to an individual's taste, which eventually homes the effectiveness of the campaigns.

Aim:

This research has a goal of finding out whether existing customers will need to renew their data packs, increasing the precision of the marketing you do.

Goal:

- Classification: Construct a model for different customers that are dividing into users and non-users of data packs.
- Insights: Reveal hidden patterns for marketing plans to boost subscription rates delivering top performance.

2. Analytical approach

a. Exploratory data Analysis

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Family	INPUT	4.411843	2.025999	38471	2037	1	4	8	0.114249	-0.98728
age	INPUT	43.97555	10.64338	38486	2022	21	42	98	0.679244	0.299916
balance	INPUT	1359.521	3088.652	40508	0	-8032	437	102114	8.521263	144.9629
campaign	INPUT	2.768243	3.09229	40508	0	1	2	58	4.810385	36.89291
duration	INPUT	264.736	255.578	39335	1173	7	188	4925	3.114605	17.86428
last_day	INPUT	15.81883	8.321429	40508	0	1	16	31	0.089555	-1.05812
passedays	INPUT	40.21848	100.0614	40508	0	-1	-1	871	2.611121	6.895413
previous	INPUT	0.581144	2.333588	40508	0	0	0	275	44.18941	4764.145

Interpretation

The dataset reveals:

Average family size: 4.41, moderately varied.

Average age: 44 years, skewed towards younger individuals.

Account balance: Highly variable, skewed towards lower balances, with outliers.

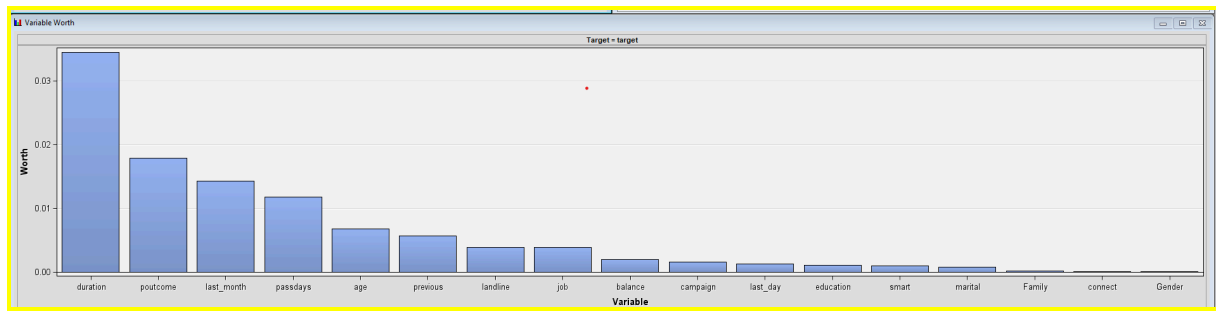
Campaign contacts: Average of 2.77, with variability and some campaigns having extensive outreach.

Call duration: Average of 264.74 seconds, highly right-skewed.

Contact frequency: Peaks mid-month, somewhat flat distribution.

Days since last contact: Highly variable, skewed towards longer intervals.

Previous contacts: Relatively low average, highly variable, and skewed towards lower values, featuring outliers.



Interpretation

The variable duration has the highest variable worth, followed by p outcome signifying it as the most important variable for prediction. Gender does not play any in prediction of whether the person will but a data pack or not

Methodology:

Concerning skewness, we'll solve the issue through transformations. Skewed variables: "education," "job," "last_month," and "marital." It has been found that classes are imbalanced in the "target" vector. The Max Normal transformation is chosen as the method. To solve the issue of missing variables we will be using impute method

b. Data Preparation

Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	Unique
Source	Imputed Variables
Role	Input
Report	
Validation and Test Data	No
Distribution of Missing	No
Status	

Data imputation:

Using the count-based method, we can impute missing values in categorical variables by substituting the most frequent category inside each corresponding variable. To put it simply, we

1. Find categorical variables that have missing values.
2. Modify the default technique for counting variables with missing values.

Imputation Summary								
Number Of Observations								
Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Family	MEAN	IMP_Family	M_Family	4.4140946234	INPUT	INTERVAL		1040
PWR_duration	MEAN	IMP_PWR_duration	M_PWR_duration	0.4444236762	INPUT	INTERVAL	Transformed duration	590
SQRT_age	MEAN	IMP_SQRT_age	M_SQRT_age	0.5317614705	INPUT	INTERVAL	Transformed age	1021
marital	COUNT	IMP_marital	M_marital	married	INPUT	NOMINAL		1040

Variable Distribution Training Data			
Obs	Number of Missing for TRAIN	Number of Variables	Percent of Variables
1	1040	2	50
2	1021	1	25
3	590	1	25

Interpretation and Results In the dataset, two imputation methods are employed: Absolute and Relative COUNT. Among interval variables that compose "Family," "PWR_duration," and "SQRT_age" categories, imputation is utilized for the values introducing as the missing ones, to fill the voids with the averages of the rest of data points. In addition, this practice allows for the data to be an accurate statistical representative of the population. NIL is the impute technique used for the nominal variable of marital and the value missing for "marital" is replaced by the most frequent category of "married." This technique is usually applied to keep the distribution of categorical data in thought.

Data Transformation

Max normalization, which is default for transforming variables, is used to normalize between 0 and 1 values by dividing each by the maximum value of that feature. This technique guarantees a symmetrical approach for features, aids in algorithm performance, allows optimization convergence, assists in feature comprehension and adds proportional differences to data points, hence making it reliable when it comes to data preprocessing for analysis and modeling tasks.

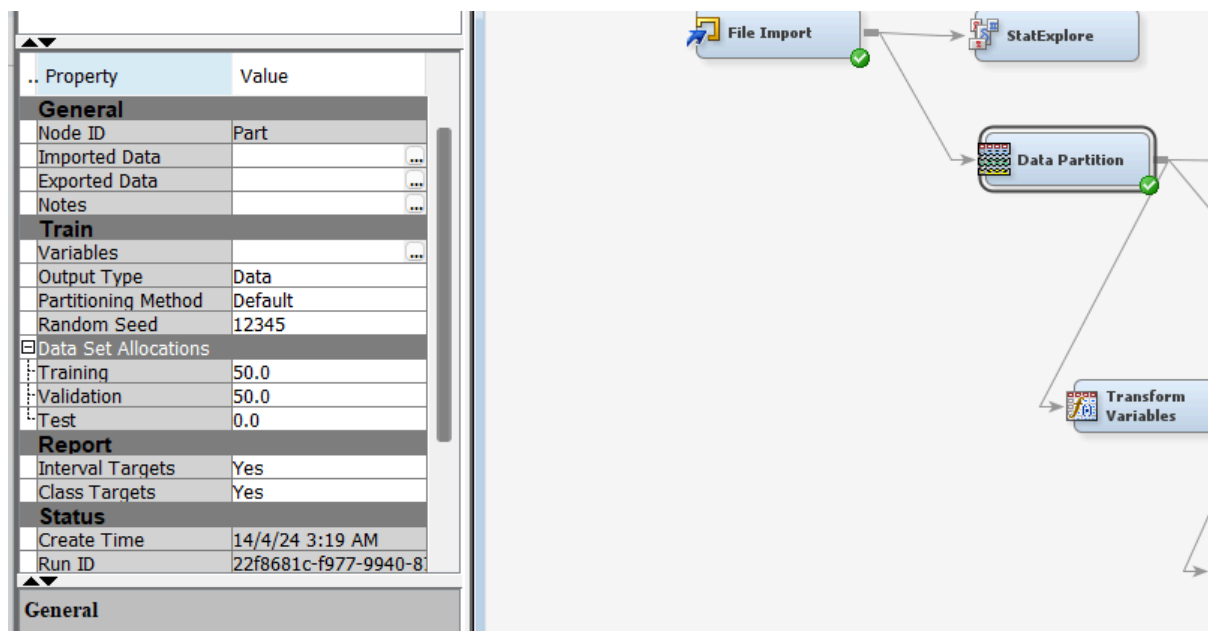
Data Partition

- **Balanced Split:** So that the distributions of the target variable and predictor of features are the same, you can partition these data into 50% for training and another 50% for testing. This will enable us to avoid such biases which could end up being caused by improper train test splits.
- **Fair Evaluation:** Overall, you'll get a more impartial sense of the model by applying the 50-50 split to evaluate its performance on a subset of data. With this, the model can efficiently appraise how well it generalizes over unseen data because the divided sample size and the composition of training and test datasets are similar.
- **Sufficient Training Data:** The training set half of data steering the model's performance to learn the patterns and relationships in the data. Hence, this avoiding the issue of underfitting by the model as the model becomes stagnant in capturing such important patterns because there are not enough training examples about to make it.
- **Robustness:** Unbalancing the data can lead to biased results, hence the reason why we equally split the data so that the model's performance metrics can be resilient and

reliable. It brings a certain level of uniformity in evaluation performance as imbalanced sampling or small test sets might introduce such inconsistencies.

- **Statistical Confidence:** As you are specifically working with a bigger test set (50% of data), you can have more certainty in the reported performance metrics such as precision, recall, accuracy, etc. This is especially important when taking decisions about the model performance for actual-world applications.
- **Consistency:** Utilizing a uniform 50-50 distribution over consecutive works or interventions is a simple measure that makes it easier to compare models' performance. It applies uniform evaluation methods framework and makes it possible to highlight either developing or turning points.

Conclusively, what SAS Miner does is to widen the data set just enough for the model to learn while keeping it big enough for a thorough assessment of model performance on new data. While the partition size is chosen for the specific context of the data set and the purpose of the analysis, the discussion above presents some general considerations that can assist with the decision making process.



Models:

2-branch tree:

- Used 2-branch decision tree as it is the basic and simplest form of decision tree. It consists of only one split, dividing the data into two branches based on a single feature and a threshold value.
- While decision stumps may seem overly simplistic compared to more complex decision trees, they can still be useful in certain scenarios for the following reasons.

- Also updated assessment measure - Average Square Error.

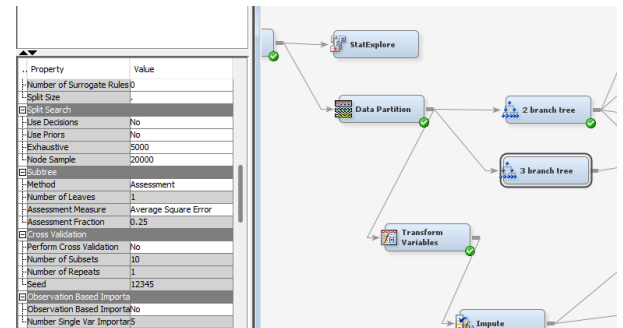
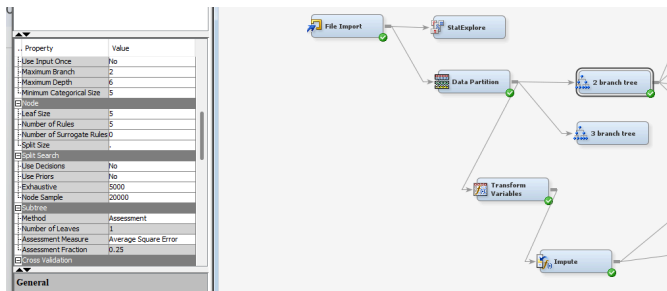
3-branch tree:

- A 3-branch decision tree, also known as a ternary decision tree, extends the concept of a decision stump by allowing for two splits instead of just one. Each split divides the data into three branches instead of two.

- While decision stumps may seem overly simplistic compared to more complex decision trees, they can

still be useful in certain scenarios for the following reasons.

- Also updated assessment measure - Average Square Error.



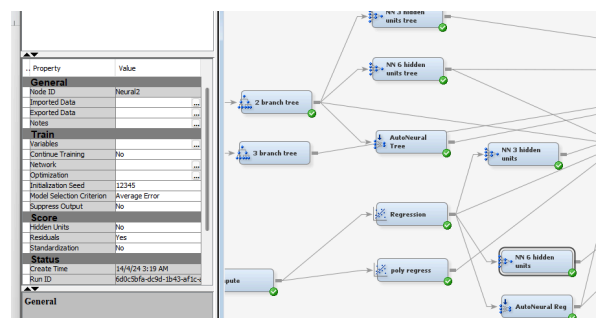
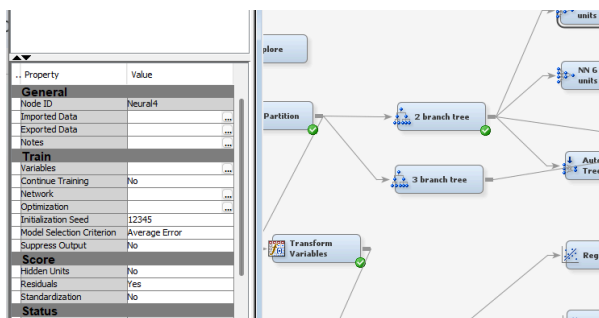
Neural Network:

Using a neural network with 3 or 6 hidden units is beneficial because:

- It captures complex relationships between features and targets.
- It automatically learns interactions between features.
- It adapts to different data types and tasks.
- It's robust to noise and irrelevant features.

Applying a neural network after regression:

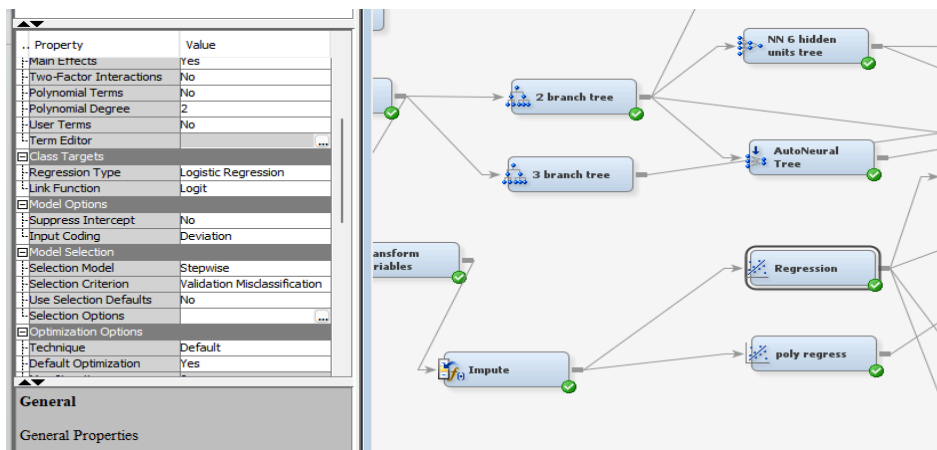
- Captures non-linear relationships missed by regression.
- Improves predictive performance by combining strengths.
- Acts as an ensemble method, enhancing overall accuracy.



Regression & PolyRegression:

1. Regression: It's used when the relationship between the independent variables and the dependent variable is linear. It predicts a continuous outcome based on the input features. It's straightforward and interpretable.
2. Polynomial Regression: It's an extension of regression but allows for non-linear relationships by adding polynomial terms (e.g., x^2 , x^3) to the model. It can capture more complex patterns in the data but may suffer from overfitting if the degree of the polynomial is too high. We select 'Yes' in Polynomial terms and 'Yes' in two-factor interactions.

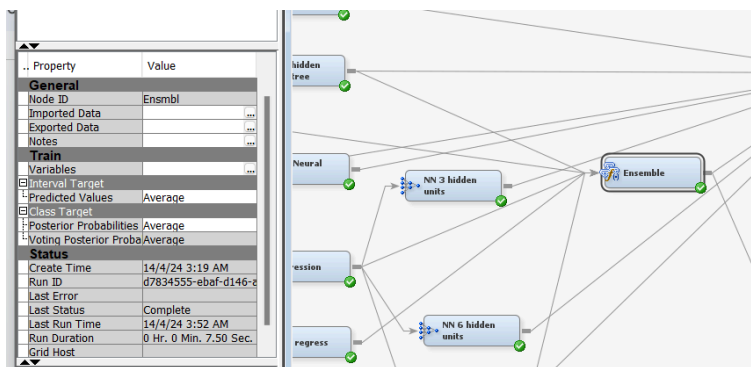
3. We have only used Validation misclassification as regression does not provide us with the option of Average Square Error or Mean Square Error.



Ensemble:

Ensemble methods in SAS Miner projects are popular because they:

- Making models more predictive by stacking them.
- Avoid overfitting the situation and increase the problemade of assistance.
- Give better results with relationships and high dimensional data, rather than be incapable of.
- Build resilience through conciseness.
- Scaling for dealing with large and in parallel is to be taken into consideration.
- Be directly interpretable by a technique like averaging models / voting.



Gradient Boosting:

Gradient boosting stands out in SAS Miner projects because:

1. It is a repetitive refinement consisting of the mistakes made by the preceding models resulting in highly precise prognosis.
2. It is able to handle with complicated data associations that makes it useful for handling diverse datasets
3. The model's behavior is made robust to overfitting by varying the model complexity. This also generally brings forth good generalization to new data.

- Such a model performs the two simultaneous tasks, not only recognizing, but also discovering which features play the greatest role in the result prediction.
- It is irrelevant whether the data is numeric or categorical as the adaptive capability of gradient boosting algorithm excels in its function and extends its flexibility.
- Using this gradient boosting technique comes as an integrated feature of SAS Miner, which, at the same time, occurs through a straightforward and effective process of data analysis for analysts.

Property	Value
Interval Bins	100
Missing Values	Largest branch
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate F0	
Split Size	
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Average Square Error
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Imputation	No
Number Single Var Imputation	5

Property	Value
Train	
Variables	
Series Options	
N Iterations	200
Seed	12345
Shrinkage	0.1
Train Proportion	50
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Split	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Largest branch
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate F0	

Results

Model Comparison

Results - Node: Model Comparison Diagram pmnnwalest

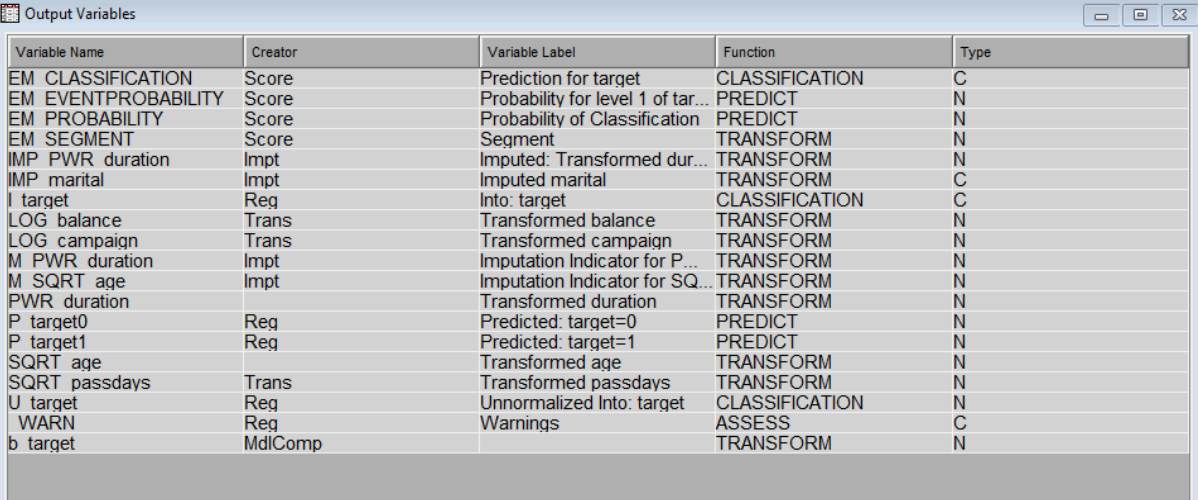
File Edit View Window

File Statistics

The charts conclude the outcome of models like Boosting, Neural Networks, and Nearest Neighbor. For each model in the table the metrics, such as AUC, and error rate are shown in detail. AUC stands for the area under the curve and is a quantification of how well good instances and bad instances are delimited by a model. An AUC of 1 would mean not helping with the problem at all, while an AUC of 0.5 would help with the problem some of the time, and an AUC of 1 would help with the problem all of the time. Misclassification rate is the number of cases the model has designated wrong over the scope of the dataset. A reduced disagreeing error indicates an excellent performance.

This evidence makes Gradient Boosting being the top performer and best amongst all the models.

Output Variables



Variable Name	Creator	Variable Label	Function	Type
EM CLASSIFICATION	Score	Prediction for target	CLASSIFICATION	C
EM EVENTPROBABILITY	Score	Probability for level 1 of tar...	PREDICT	N
EM PROBABILITY	Score	Probability of Classification	PREDICT	N
EM SEGMENT	Score	Segment	TRANSFORM	N
IMP PWR duration	Impt	Imputed: Transformed dur...	TRANSFORM	N
IMP marital	Impt	Imputed marital	TRANSFORM	C
I target	Req	Into: target	CLASSIFICATION	C
LOG balance	Trans	Transformed balance	TRANSFORM	N
LOG campaign	Trans	Transformed campaign	TRANSFORM	N
M PWR duration	Impt	Imputation Indicator for P...	TRANSFORM	N
M SQRT age	Impt	Imputation Indicator for SQ...	TRANSFORM	N
PWR duration		Transformed duration	TRANSFORM	N
P target0	Req	Predicted: target=0	PREDICT	N
P target1	Req	Predicted: target=1	PREDICT	N
SQRT age		Transformed age	TRANSFORM	N
SQRT passdays	Trans	Transformed passdays	TRANSFORM	N
U target	Req	Unnormalized Into: target	CLASSIFICATION	N
WARN	Req	Warnings	ASSESS	C
b target	MdlComp		TRANSFORM	N

The table features the name of variables, their labels, functions, and types. The variables are listed as per their importance, which is decreasing, from first to fifth, while the scores are provided in the same way, only with decreasing importance, from first to last.

Here are some of the things that can be found in the table: Here are some of the things that can be found in the table:

Variable Name: This is the name of the variable which is mined from the data itself or machine learning procedures. This, for instance, is the identifier for 'EM CLASSIFICATION', which could very well be the output of a classificatory model.

Variable Label: The purpose of the labels is that they need to be easily understandable by humans. Take, for example, the case of variables labels: "EM CLASSIFICATION"—> "Prediction for target".

Function: This specifies the utilization of the provided function in order to build the variable. The sub-form for "EM CLASSIFY" reads as "CLASSIFY", for instance.

Type: This just indicates the type of the data that the variable contains. One illustration is the way that "EM TYPE" is "C", which could indicate there is a closed data type, or categorical data.

Conclusion

The custom data analysis from the case study of a telecom company contains gold insights for the data pack campaigns that help increase the revenue. Our approach that relies on a combination of decision tree, neural network, regression estimation and ensemble methods took us to the top of customer packs data usage learning curve.

An investigation showed that call duration is the prevalence factor for clients who renew data packs and a call center hosted campaigns, to be precise. Infact what was greatly striking was the statistically inconsequential link between gender and the data packs usage.

Our data imputation methods and normalization techniques addressed the quality data concern and addition ensure a robust model. This equally partitioned data added the credibility of the research's conclusion to a certain extent.

The Gradient Boosting model emerged as the most performant model, achieving superior accuracy in predicting data pack users. This model not only delivers precise predictions but also highlights the key features driving those predictions, enabling targeted marketing strategies.

These insights empower the telecommunications company to:

Personalize data pack offerings: Tailor marketing campaigns based on factors like call duration and past campaign response, maximizing campaign effectiveness.

Focus resources: A locate marketing efforts towards customer segments most likely to benefit from data packs, optimizing campaign ROI.

Predict customer behavior: Anticipate customer data pack needs, allowing for proactive outreach and retention strategies.

By implementing these data-driven recommendations, the telecommunications company can significantly enhance customer engagement and drive data pack subscriptions. Future research avenues could involve exploring additional customer data points and incorporating time-series analysis to further refine customer segmentation and campaign targeting.

Recommendations

- Develop targeted marketing campaigns: Benefit from the fact of gradient deduction for segmenting customers according to call duration, past campaign response and later on other factors. Develop tailored marketing messages that will connect to each client group and bring the most appropriate characteristics of data packages concerning their behavior patterns.
- Prioritize high-value customers: Personify customers with high call length and positive call handling history in doing past campaigns. Design tailored data packs for specialized target groups and membership programs that will reward the consumers and have a positive impact on the customer life-time value.
- Optimize campaign timing: Elaborate on employing time-series study to isolate the seasonal pattern and spots of best interests for radiating promotional messaging to sales of data packs. Those act like expanding the reach as well as improving sales and conversion rate at the similar price.
- Invest in customer education: Create educational material that breaks down the benefits of plan and demonstrate the various usage scenarios that show the value of data plans. This could be very helpful in showing and stimulating readiness consequently the customers who have lower call durations to buy the product.
- Monitor and refine strategies: Make an ongoing effort to track campaign performance and customer trends. Use our AI to write for you about the given topic "Intangible Assets in the Digital Age". Our AI will write a high-quality essay for you in just 30 seconds. This allows you to focus on other aspects of your coursework or job while

still getting the information you need. Sharpen your concentrate to be more effective with the latest data in hand to keep the campaign effectiveness at the highest level.

Future Areas of Work

- Incorporate additional data points: Add more new data, like the customers' demographics, internet usage patterns, and device preferences, to make the analysis more solid. Such an effect can help to know the behavior of a customer better and better fitting a profile as a result.
- Explore churn prediction: Explore the development of models for predicting customer churn, permitting timely preventive remedies to retain loyal customers. This can be realized by scrutinizing metrics such as the customer satisfaction index, service experience data, and competitors' products,
- Investigate RFM analysis: Introducing RFM model for grouping customers according to purchase patterns (Recency, Frequency, and Monetary analysis). This is where you can get data regarding the exact same thing and find out the consumer segments that are most lucrative for you to target with data packs.
- Integrate A/B testing: Set up A/B testing to evaluate and discover the promotional creatives, messaging strategies, and campaigns channels that worked best. This process of creating based on analyzed data helps detecting of the most efficient directions that will be used in future campaigns.
- Investigate alternative models: The fact that the gradient boosting algorithm showed the best results does not leave room for other machine learning models to be explored – namely the Random Forests and Support Vector Machines (SVMs), as they might have higher predictive power within certain customer groups.

Through the implementation of these fields of activity, telecommunications enterprise would build an informationally based marketing plan, which will ensure customer engagement, receiving of the data packs and the growth of the revenue.