MLERS Data Report

Emmaculate Akoth, Faithelizabeth Mburuga, Derrick Kuria

& Collins Kemboi.

Table of Contents

Business Understanding

Introduction

MegaTelco USA is an integrated telecommunications provider in the USA. It was previously a part of the USA Posts and Telecommunications Corporation (USAPTC) which was the sole provider of both postal and telecommunication services. The company was established as a telecommunications operator in April 1989, after the split of USAPTC into the Communications Commission of USA (CCRA), the Postal Corporation of USA (POSTA) and MegaTelco USA. The company is 60 per cent owned by Helios Investment Partners, with the remaining stake held by USAns through the Government of USA.

MegaTelco USA provides integrated telecommunications solutions to individuals, Small and Medium-sized Enterprises (SMEs), Government and large corporates in the USA, drawing from a diverse solutions suite that include voice, data, mobile money as well as network services. Powered by its vast fibre optic infrastructure, it is also a major provider of wholesale, carrier-to-carrier traffic within the country and the region.

The company operates and maintains the infrastructure over which USA's various internet service providers operate. As of 2004, most internet service was provided via dial-up service. Viasat, an important USAn ISP, is a subsidiary of MegaTelco USA. It also offers mobile GSM voice and high speed internet services under the Viasat USA brand, in which it is the 3rd in market share after At & T and Verizon. In March 2018, the company resumed a mobile-money service that it had dropped in 2017.

Referred to as Mega-kash, the service is a direct competitor to theV-transfer service, offered by market-leader Verizon.

Problem Statement

MegaTelco is a telecommunication company that aims to provide quality services to its subscribers and foremost aims to become a leader in the industry which will result in increased revenue to drive social and economic impact in the Country through retention of its clients.

Churning is expensive to the company that loses the customer as attracting new customers is more expensive than retaining old ones. The company has contacted us to build models to predict whether a customer will change from one Telecom company to another(churn).

The models will therefore provide the company with key insights that will assist in devising action plans on how to prevent or reduce churning. This information will be passed to the Marketing department to create special offers to the identified subscribers with the aim of retaining them.

Objective

The objective of this analysis is to develop different models using Logistic regression, Decision Trees, SVMs, KNNs and Naive Bayes that will successfully be deployed to predict churning, that is whether a customer will change from one Telecom company to another.

The secondary objective is uncovering insights on how to prevent churning through determining key factors that would make a subscriber churn to other providers.

Data Understanding

The data that will be used for our analysis was from MegaTelco and has information on number of voicemail messages, total calls per day, number of customer service-related calls and other usage information. The data was downloaded in csv form which makes it easier for us to load and conduct further analysis.

The data used has 18 features and 5000 entries. The features of the data include:

Feature	**Explanation**	**Feature**	**Explanation**
Churn	Whether the customer churned, [Yes, No]	totalintlminutes	Total International minutes
accountlength	Account character length	totalintlcalls	Total International calls
internationalplan	Subscription to International Plan, [Yes, No]	totalintlcharge	Total International charge
voicemail	Subscription to voicemail, [Yes, No]	numbercustomerservicecalls	Number of calls made to Customer Service
numbervmailmessages	Number of voicemail messages		

totaldaymin utes	Total day minutes		
totaldaycall s	Total day calls		
Totaldaych arge	Total day charge		
totalevemin utes	Total evening minutes		
totalevecall s	Total evening calls		
totalevech arge	Total evening charge		
totalnightmi nutes	Total night minutes		
totalnightca lls	Total night calls		
totalnightch arge	Total night charge		
Date	Date of record		

Data Preparation

Data preparation included cleaning the dataset and selecting the variables that would be used for Univariate analysis, Bivariate analysis, and Multivariate analysis.

Below are the steps we took to clean our data;

Step 1:

Data cleaning was the first data preparation step taken. Data cleaning involved checking for missing values and duplicate values. In our case we did not have any missing or duplicated data.

Step 2:

The next step was to check for outliers. We plot a boxplot of all our numerical data in the dataset. It was visible there was a presence of outliers. Through the use of Z-score we dropped all numerical records with a score of above 4 to retain the integrity of our data. A total of 48 records were therefore dropped.

Step 3:

To ensure that the data was ready for univariate analysis, the dataset was divided into numerical and categorical variables. This allowed for numerical and categorical variables to be analyzed and explored effectively. By generating the frequency table for the categorical variable, it was possible to plot frequency distributions for specific variables. A frequency distribution table was also developed from the numerical variables.

Only the numeric variables were used in the bivariate analysis. The data frame that was separated for univariate analysis was used for bivariate analysis. Similarly, only the numeric variables were used to plot the correlation matrix and a double bar graph of our categorical data churn against numerical variables.

Step 4:

For multivariate analysis, both numeric and non-numeric variables were used. The categorical variables used were encoded to make sure that they could be used for dimensionality reduction.

Data Analysis

We analyzed our data in the levels to uncover key insights and also to select key features as shown below.

Univariate Analysis

Univariate analysis provides summary statistics for each field in the dataset that is on each variable. On our univariate analysis we plotted a frequency distribution on each variable.

From our observation it was clear the variables total_intl_charge, total_intl_minutes,total_night_charge, total_night_calls, total_night_minutes, total_eve_charge, total_eve_calls, total_eve_minutes, total_day_charge, total_day_calls, total_day_minutes and account_length are normally distributed. It was also observed that the Total international calls and number of voicemail messages were skewed on the right and lastly a number of customer service calls had a bimodal distribution.

Bivariate Analysis

Bivariate analysis is performed to investigate the relationship between two variables in a dataset. Bivariate analysis was conducted on numeric variables.

The bivariate analysis included a double bar graph and a correlation matrix. Below are the key observations we were able to conclude from our bivariate analysis:

Observation

1. Of the people with no voicemail plan, a bigger proportion are not likely to churn (switch from one company to another)

2. A bigger proportion of people with no international plan are likely to not churn.

3. The variables that have the least correlation to churn are:
    - voicemail plan
    - number of voicemail messages
    - total eve calls
    - total night calls
    - total international calls
4. The variables that have the highest correlation to churn are:
    - Number of customer service calls
    - Total day charge
    - Total day minutes
    - International charge

Multivariate Analysis

Linear Discriminant analysis was used to implement the multivariate analysis as a dimension reduction technique. In the multivariate analysis, the churn variable was the dependent variable while the independent variables were the rest of the variables.

We were able to observe that only 10 features were suitable to determine if a customer would churn or not. The features are;

1. total_day_minutes
2. international_plan
3. total_night_charge
4. number_customerservice_calls
5. total_intl_charge
6. total_eve_minutes
7. number_vmail_messages
8. total_intl_minutes
9. total_day_calls
10. account_length

Modeling

The data was modelled using classification techniques and Logistic Regression. Classification is a process of predicting the class of given data points. This type of predictive modelling allows data scientists to approximate a mapping function from input variables to output variables. The classification techniques used are Decision Trees(Random Forest Classifier), SVMs, KNNs and Naive Bayes.

Logistic Regression

Logistic regression, also called a logit model, is used to model dichotomous outcome variables. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables.

Our Logit model was 75% accurate, which did not meet our threshold on our metrics of success gauge.

Decision Trees

A decision tree is a modeling technique that uses a tree-like model of decisions and their possible consequences. We used it to create a model that was able to help us identify customers who are likely to churn upon expiration of their contracts.

We used Random Forest Classifier and Gradient Boosting Classifier to model our data. Our Random Forest Classifier model was 93% accurate therefore able to predict accurately while Gradient Boosting Classifier was 99% accurate which performed better than the previous models.

SVMs

The Support Vector Machine is a supervised learning algorithm that we will use in classification on whether a customer is likely to churn. The algorithm will be able to classify if a customer will churn or not based on that data provided. This occurs when the SVM algorithm finds a hyperplane in the number of features dimension space that distinctly classifies the data points.

Our SVM model was very accurate at 100%.

KNNs

K Nearest Neighbours Classification is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

The KNN model developed was 96% accurate.

Naive Bayes

Naive Bayes classifier looks at independent features separately to determine how these features contribute to the probability of getting the dependent variable. The dependent variable that was being predicted using Naive Bayes is the churn. The features that were analysed to determine the churn are the independent variables from our dataset.

Our Naives Bayes model has an accuracy of 96%. This is high enough and can make accurate enough predictions.

Evaluation

From all our six models it was noted that the SVM model had the highest accuracy of 100%, followed by the Gradient Boosting Classifier model at 99% accuracy. However, the model best placed to determine which customer will switch from one company to another is the KNN model with an accuracy of 96% which is sufficient to do the predictions.

These accuracies though high, are prone to overfitting. The model best placed to determine which customers will switch from one company to another (churn) is the KNN model.

Recommendations

1. **Ask for feedback** : A loyal customer is always willing to give feedback on how to make the product better. .Ask for feedback using surveys and calls. For all answers surveys, giving an incentive would be a good act of appreciation to the customer. This increases loyalty to the product.
2. **Listen to the Customer** : This Is not only taking grievances by customers but by also ensuring that the customer care department is up and running. Also the customer care agents should be receptive and accommodating to the different characters of customers
3. **Stay Competitive** : To avoid customer churn, loyal customers are always anticipating changes, new advertisements , better rates and good performance compared to competitors. Hence, staying competitive shows continued vigorous activity and focus towards the end product to the customer

Deployment

The models will be used to predict customers who are more likely to churn. The information will be passed to the marketing department of the company to offer the identified customers special offers in order to retain them.

Reference

Link to the Python Notebook : https://bit.ly/2NPSDus

Link to the dataset : https://data.world/datasets/classification