Determine Churn Rate

IST 707: Applied Machine Learning

By-
Ashitosh Gupta
Nidhee Patil
Chandra Shekar Manthena
Zack Dareshori
IST 707

Contents

Introduction:

Our business problem: Telecom Churn Analysis

The background of our business problem:

Churn rate is the rate at which customers stop doing business with a company over a given period. Churn may also apply to the number of subscribers who cancel or don't renew a subscription. The higher your churn rate, the more customers stop buying from your business. The lower your churn rate, the more customers you retain. Typically, the lower your churn rate, the better.

Understanding your customer churn is essential to evaluating the effectiveness of your marketing efforts and the overall satisfaction of your customers. It's also easier and cheaper to keep customers you already have versus acquiring new ones. Due to the popularity of subscription business models, it's critical for many businesses to understand where, how, and why their customers may be churning.

Churn rate of a company can be calculated by subtracting the no of customers left with no of customer at the beginning of a specific period, then divide it with no of customers at the beginning of the period and multiply the whole with 100.

After brain storming, we decided that the churn rate can be prevented by the following procedures. By understanding why customer churn (to find patterns), and by predicting whether a person or going to be a churn or not (A sign).

Data Description:

This dataset is collection of customers details of an unnamed company from the telecom industry. This dataset is basically describing about the customer churn (True/False). There are 20 features with 3333 instances.

```{r}
str(prep_data)
```

```
'data.frame':   3333 obs. of  20 variables:
 $ State                 : chr  "KS" "OH" "NJ" "OH" ...
 $ Account.length        : int  128 107 137 84 75 118 121 147 117 141 ...
 $ Area.code             : int  415 415 415 408 415 510 510 415 408 415 ...
 $ International.plan    : chr  "No" "No" "No" "Yes" ...
 $ Voice.mail.plan       : chr  "Yes" "Yes" "No" "No" ...
 $ Number.vmail.messages : int  25 26 0 0 0 0 24 0 0 37 ...
 $ Total.day.minutes     : num  265 162 243 299 167 ...
 $ Total.day.calls       : int  110 123 114 71 113 98 88 79 97 84 ...
 $ Total.day.charge      : num  45.1 27.5 41.4 50.9 28.3 ...
 $ Total.eve.minutes     : num  197.4 195.5 121.2 61.9 148.3 ...
 $ Total.eve.calls       : int  99 103 110 88 122 101 108 94 80 111 ...
 $ Total.eve.charge      : num  16.78 16.62 10.3 5.26 12.61 ...
 $ Total.night.minutes   : num  245 254 163 197 187 ...
 $ Total.night.calls     : int  91 103 104 89 121 118 118 96 90 97 ...
 $ Total.night.charge    : num  11.01 11.45 7.32 8.86 8.41 ...
 $ Total.intl.minutes    : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
 $ Total.intl.calls      : int  3 3 5 7 3 6 7 6 4 5 ...
 $ Total.intl.charge     : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
 $ Customer.service.calls: int  1 1 0 2 3 0 3 0 1 0 ...
 $ Churn                 : Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...
```

Goals for our project:

1. understanding why a customer is a churn.

2. by predicting whether a person or going to be a churn or not?

Methods:

Our group members have decided on performing Exploratory Data Analysis, Association Rule Mining, Clustering, and few Machine Learning Models on the data to achieve our goals of the project.

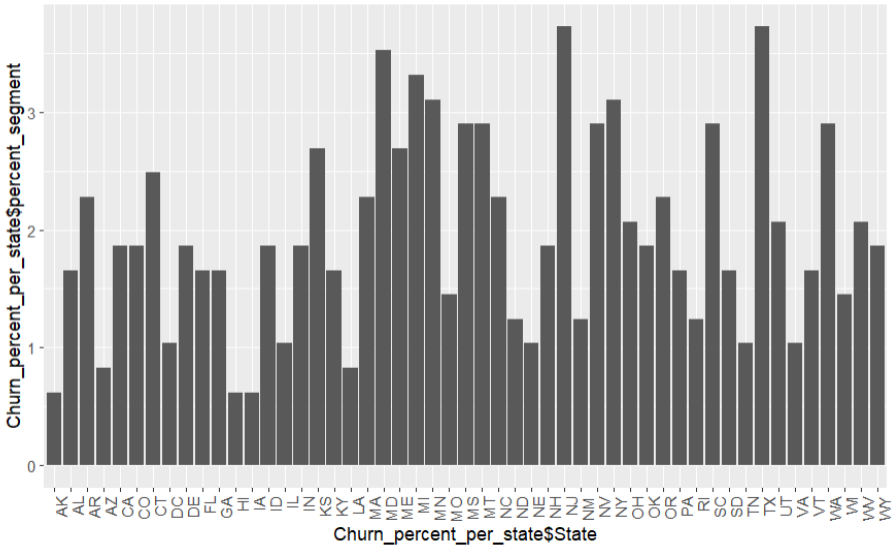
Exploratory Data Analysis:

Data cleaning and pre-processing:

After checking the dataset, we found that some features state, international plan, voice mail plan are not categorical variables. So, converted them in categorical variable using as.factors(). Since we haven't found any null values, so we haven't performed any interpolation methods. But for association rules we have discretize the numerical variables into categorical variables, so that we can identify the patterns.

1.State:

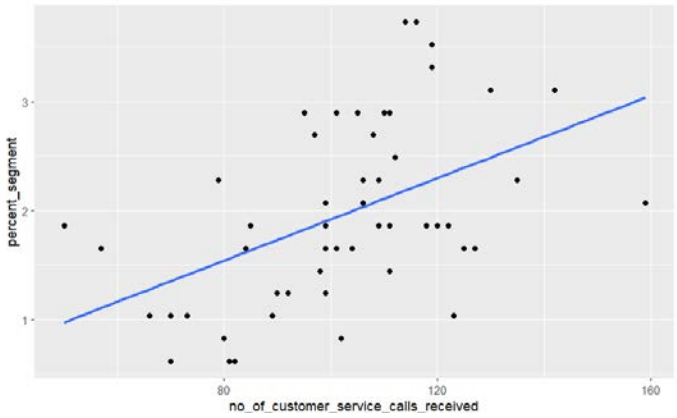
Found that 33% churn rate occurred in the telecom company is from 20% of the states, i.e 11 states causing around 5% churn rate in overall 14.5%.



NJ, TX, MD: has more than 0.5 churn rate in overall (14.5%)

MI, MN, NY, MS, MT, NV, SC, WA: has more than 0.4 churn rate in overall (14.5%)

We have observed a correlation in between the churn_percent_per_state with average_customer_servive_call_per_state. Can't prove why the correlation is present with currently available data, need the customer geographical location, signal tower location and context of customer service calls.



Assumptions:

1. Low signal

2. Competitors' domination (offers, better price)

2.Charge:

There are different charges for different time periods of a day. Telecom company is charging differently in day, evening, night, and international calls.

We wanted to see, if there is any price difference for the people how has Churn=True with people how has Churn=False.

Used few statistical inference techniques to study the difference and interpret them:

T-test and Bayesian analysis:

t-test:

The p-values of all types were seen as less than the 0.05, which states they are statistically significant, saying there is difference in between the charge paid by the people who are churn and in between the people who are not churn. The Lower and Upper bound gives the variable in the difference.

Call type	t-value	p-value	Lower-bound	Upper-bound	x-mean (True)	y-mean (False)
Day	9.6845	<0.05	4.30	6.480	35.175	29.78
Evening	5.272	<0.05	0.712	1.559	18.054	16.918
Night	2.171	<0.05	0.021	0.4369	9.2355	9.0060
International Calls	3.9399	<0.05	0.0733	0.2189	2.8895	2.7434

Bayesian analysis:

The Bayesian analysis gives the probability distribution, which says that the means are most probably observed difference in between the charge paid by churn customer and non-churn customer.

Call Type	Mean	Lower-Bound	Upper-Bound	Overlapping Zero	Data below Zero
Day	5.4	4.31	6.51	NO	0
Evening	1.13	0.706	1.55	NO	0
Night	0.23	0.017	0.434	YES	1.7%
International Calls	0136	0.064	0.209	NO	0

3.International.plan:

We have observed that the average international talk time of the customers are almost same regardless of their status on internatioanal plan and customer churn.

International.plan	Churn	Average.international talktime
Yes	True	11.78
Yes	False	9.778
No	True	10.27
No	False	10.19

We have performed t-test and Bayesian analysis on they charge of customer with and without international plan, and observed that

t-test:

	t-value	p-value	Lower bound	Upper bound	x-mean	y-mean	
International charge with respective to the international plan	2.7254	0.006704	0.03250	0.20075	2.8699	2.753279	

So, the p-value is significant saying there is a price difference but it’s in between 0.0325 and 0.2

Bayesian Analysis:

	Mean	Lower-Bound	Upper-Bound	Overlapping Zero	Data below Zero
International charge with respective to the international plan	0.111	0.0299	0.195	Yes	0.4%

The Bayesian analysis gives the probability distribution, which says that the means are most probably observed difference in between the charge paid by customers having international plan and customer without international plan.

Feature selection:

Generalized Linear Model:

Using glm, I have checked for the "Which features has a significant impact on the churn?". The result has showed that customer.service.calls, international.plan, voice.mail.plan, The prediction has not yet done at this point with this model.

```
Call:
glm(formula = Churn ~ . - State, family = binomial(link = "logit"),
    data = prep_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1595  -0.5127  -0.3402  -0.1957   3.2500

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -8.4416761  0.9257897  -9.118  < 2e-16 ***
Account.length         0.0008311  0.0013919   0.597 0.550421
Area.code             -0.0004771  0.0013134  -0.363 0.716425
International.planYes  2.0456277  0.1457194  14.038  < 2e-16 ***
Voice.mail.planYes    -2.0250756  0.5740856  -3.527 0.000420 ***
Number.vmail.messages  0.0358748  0.0180108   1.992 0.046388 *
Total.day.minutes     -0.2566903  3.2747702  -0.078 0.937522
Total.day.calls        0.0032016  0.0027613   1.159 0.246278
Total.day.charge       1.5861483 19.2634096   0.082 0.934376
Total.eve.minutes      0.7916394  1.6374536   0.483 0.628771
Total.eve.calls        0.0010531  0.0027829   0.378 0.705116
Total.eve.charge      -9.2280330 19.2640598  -0.479 0.631918
Total.night.minutes   -0.1126807  0.8772554  -0.128 0.897795
Total.night.calls      0.0007167  0.0028425   0.252 0.800943
Total.night.charge     2.5859921 19.4938976   0.133 0.894465
Total.intl.minutes    -4.2873631  5.3027500  -0.809 0.418793
Total.intl.calls      -0.0928850  0.0250540  -3.707 0.000209 ***
Total.intl.charge     16.2026936 19.6390321   0.825 0.409357
Customer.service.calls 0.5139115  0.0392868  13.081  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2158.6  on 3314  degrees of freedom
AIC: 2196.6

Number of Fisher Scoring iterations: 6
```

Rpart model:

This is like decision tree model, which shows the distributions of the dependent variable with respect to predictive variables. Currently facing few errors in visualizing the tree distribution in R programming.

	Overall <dbl>
Total.intl.minutes	100.000000
Total.day.minutes	88.487866
International.planYes	86.585689
Customer.service.callshigh1	70.123268
Total.intl.calls	39.264941
Voice.mail.planYes	36.369492
Number.vmail.messages	30.367304
Total.eve.minutes	11.669889
Total.night.minutes	9.426877
Customer.service.callshigh2	5.354199

We used various methods to check for valuable features for modelling, we have performed generalized linear model, rpart model, PCA component method and

By using L1 Regularization in the logistic regression model, to obtain some important features. The features are as listed below.

Important features: ['Total day minutes', 'Total day calls', 'Total day charge',

'Total eve minutes', 'Total eve charge', 'Total night minutes',

'Total night calls', 'Total night charge', 'Total intl calls',

'Total intl charge', 'Customer service calls', 'International plan_No',

'International plan_Yes', 'Voice mail plan_No', 'Voice mail plan_Yes']

Clustering:

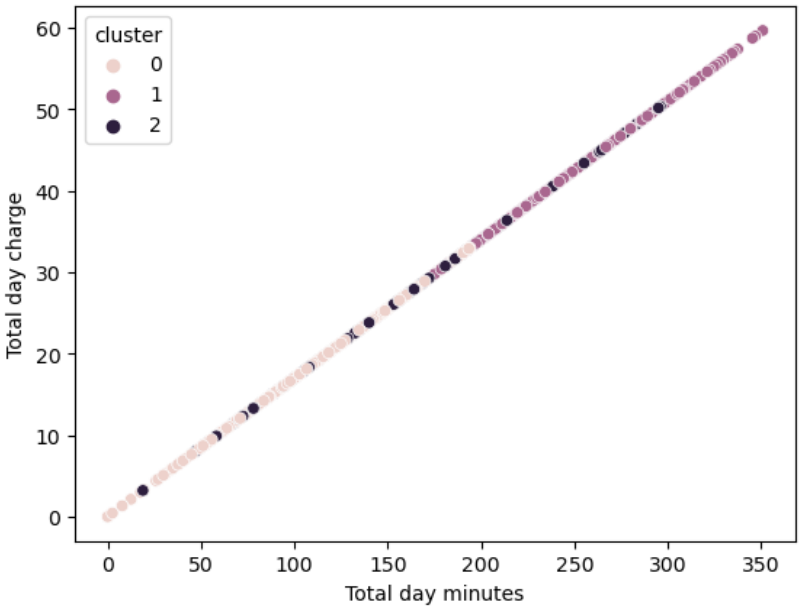
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

We will utilize clustering to identify any shared characteristics among the telecom customers and put them into groups. Then, we may observe how retention works in relation to these groups.

For this purpose, we have K-means clustering:

- K-means clustering: the K-means method finds k centroids and then assigns each data point to the closest cluster while minimizing the size of the centroids.
- To choose K, I first used the elbow method such that as the number of clusters increases, the variance (within-cluster sum of squares) decreases. The elbow at 2 or 3 clusters represents the most parsimonious balance between minimizing the number of clusters and minimizing the variance within each cluster hence we can choose a value of k to be 2 or 3.
- I also reconfirmed the K value by calculating Silhouette scores and found k=3 to have the highest score and hence decided to use k=3.
- After performing K-means clustering, I found some distinct clusters for a few features and visualized and explained below.

Example 1: The following figure shows us the distinct clusters formed among customers having high total charge and high number of total minutes they talk in day compared to another customer having low total charge and low minutes they talk in day.

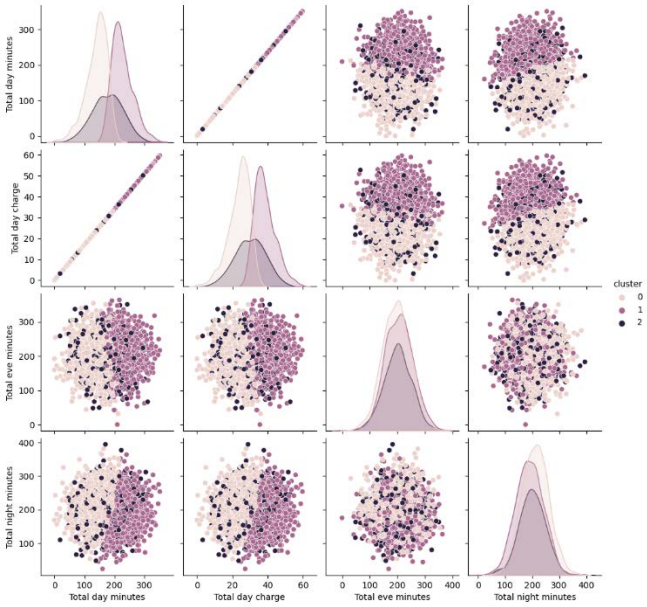


Cluster formed with Total day charge Vs Total day minutes

The figure below shows two clear clusters with interpretations as follows:

1. One group has high charges with many minutes they talk in day evening but less at night while having other factors constant.
2. Second group has less charge, talks less minutes in day and evening but talks more at night compared to the first group keeping other factors constant.

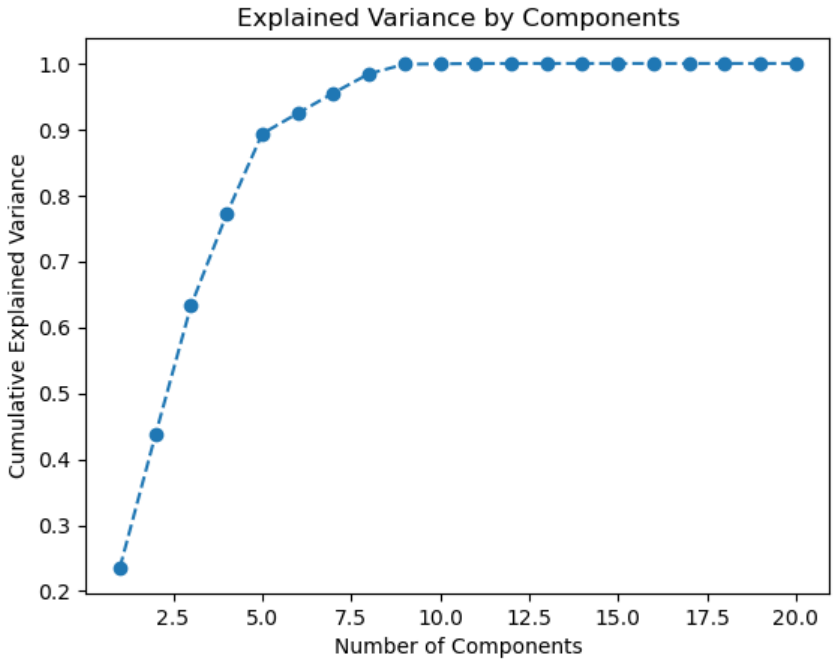
The figure below summarizes the clusters for each influential factor.

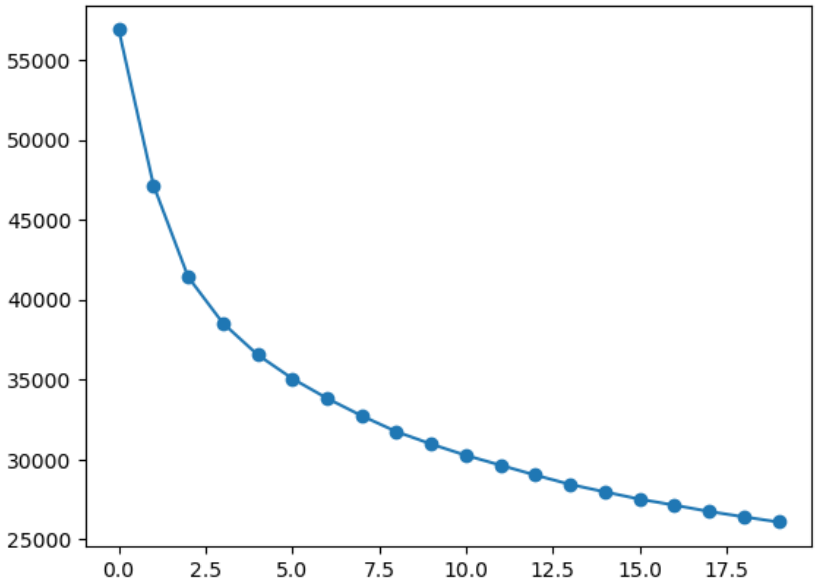


Cluster formed with different features

Principal Component Analysis:

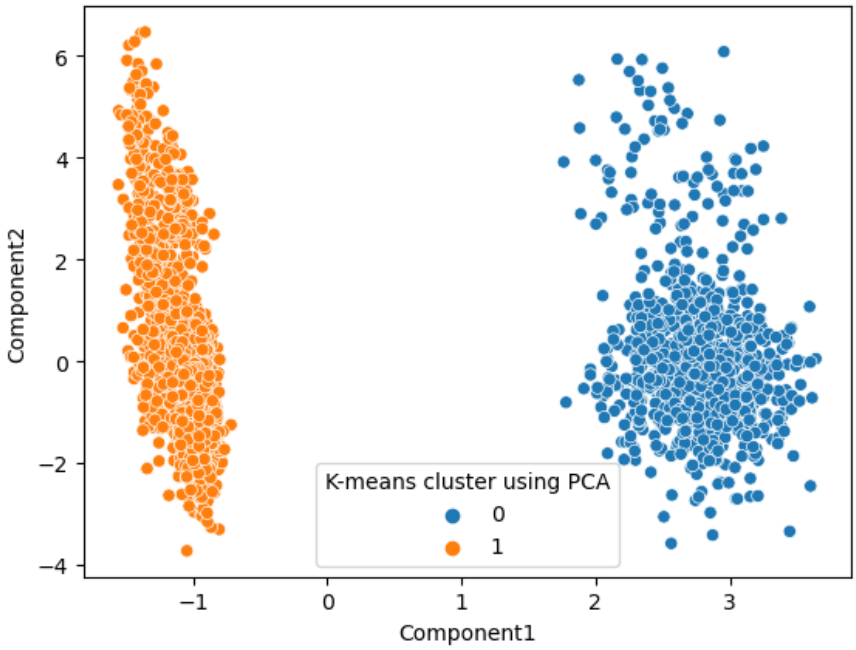
The clustering did not provide us with distinct clusters by considering individual clusters and hence there has to be multiple variables working together to influence the churn rate. And also since there are many features in the dataset we decided to use Principal Component Analysis. Principal component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. As we can see approximately 10 components cover the variance by 100%.





Required clusters using Elbow method

After forming the PCAs and performing clustering using the above steps, the result improved significantly giving a very distinct cluster as shown below.



Cluster formed with PCA1 Vs PCA2

	Component1	Component2	Component3	Component4	Component5	Component6
International plan_Yes	-0.015214	0.486519	-0.101723	0.037893	-0.002785	0.461482
International plan_No	0.015214	-0.486519	0.101723	-0.037893	0.002785	-0.461482
Total day charge	-0.022311	0.365364	0.336131	0.126767	0.266916	-0.379003
Total day minutes	-0.022310	0.365362	0.336129	0.126771	0.266918	-0.379004
Churn_num	-0.095381	0.340820	0.075877	0.031324	-0.040949	0.028278

Correlation of PCA with some of the features

Insights:

- We can observe that Churn is highly positively correlated with customer service calls, total day minutes, total day charge and if the customer has an international plan.
- Churn is highly negatively correlated with no international plans included and total international calls.

Model Creation:

Support Vector Machine:

Support vector machines are a type of supervised learning methods that work by plotting data points in high dimensional space to find the relationships between them and uses that knowledge to generate predictions for new data points.

To implement SVM, we used the NuSVC model from Scikit-Learn (sk-learn). This is a type of SVM that uses a "nu" parameter instead of C to control regularization, but it is essentially the same as a classic SVM. After adjusting the parameters using a grid search, a nu value of

0.2 was used, and a non-linear kernel was chosen, specifically the radial basis function kernel. Using 10-fold cross-validation, an F1 score of .63 was achieved on the full dataset. Since the dataset was unbalanced, we used under sampling to generate a smaller, balanced dataset. The SVM achieved an F1 score of .65 on the under sampled dataset.

Gradient Boosting Trees:

Gradient Boosting Tree (GBT) classifier is a machine learning algorithm that works by using a group of sequentially trained decision trees to arrive at a strong classification. Unlike a random forest, in which the trees are trained in parallel, each tree in a GBT classifier builds on the error of the previous tree in a sequential manner. This allows it to learn complex functions.

To implement GBT, we used the sk-learn GradientBoostingClassifier with 1,000 trees specified. Using this method, we achieved an F1 score of .827. We also achieved the same score on a balanced test set built with under sampling (we used the imblearn Random Under Sampling method with the ratio strategy for under sampling).

Another GBT was trained on data that was oversampled using Synthetic Minority Oversampling Technique (SMOTE). That model achieved an F1 score of .818 on the regular test set, and .802 on the under sampled and balanced test set.

Random Forest:

A random forest classifier is an ensemble learning method which is similar to GBT but the trees are trained in parallel, and then they "vote" on the outcome. The class which gets the most votes win. They operate on a similar premise to "wisdom of the crowd".

To implement our random forest classifier, we used the sk-learn Random Forest Classifier. For the number of trees, we achieved the best results with 100 trees. It makes sense that we used less trees than with GBT because unlike GBT, using too many trees in random forest can lead to overfitting.

In 10-fold cross-validation, we achieved an F1 score of .834 with random forest. With oversampling, that jumped to .838. On under sampled test data, the score dropped down to .79.

We also implemented recursive feature elimination (RFE) with random forest. Recursive feature elimination is a method feature elimination that works by recursively removing features that are the least relevant to the dependent variable. Using RFE, we were able to remove 11 of

the features in the dataset and achieve an F1 score of .77 with only 10 features. The features that remained are listed below:

- Total day minutes
- Total day charge
- Total eve minutes
- Total eve charge
- Total night minutes
- Total international minutes
- Total international charge
- Customer service calls
- International plan (boolean)

Naive Bayes:

- It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. Naive Because it presumes that each input variable is independent, Bayes is referred to as naïve. This is an overly optimistic assumption based on hypothetical data, yet the strategy works wonders for a variety of difficult problems.
- Using the data set we have; we aim to predict the people who churn or do not churn based on the factors affecting their attrition.
- Therefore, to precisely predict whether they will retain their plan with the same telco company. We needed to under sample and some pre-processing steps. To make the data a well-balanced dataset. This was made possible by python and orange 3. In orange 3, the data sampler along with the concatenate widget has helped to reduce the dataset to a well-balanced 51:49 ratio.
- Finally, after performing the under sampling the model was runned to give us a prediction F1 score of 85.6%, Precision of 85.6%. This is a pretty good score, to know if the customers would stay or rather leave the plan.

Decision Tree:

- The non-parametric supervised learning approach used for classification and regression applications is the decision tree. The objective is to learn straightforward decision rules derived from the data features to build a model that predicts the value of a target variable. Therefore, to predict or classify the people leaving the plan or organization or rather continuing with the plan.
- Based on the variables influencing retention, our goal is to anticipate the individuals that churn or do not churn.
- Here, the pre-processing steps are used in orange 3. We have utilized the select column widget to set our target variable as churn. Then further on used the select rows to set a condition for churn=0 and pass it through the data sampler to sample 487 rows of data, as the dataset as a very large number of data rows with churn equal =1. These steps were performed to balance out the dataset to yield better results. The final data was linked by using the concatenate widget and we receive a well-balanced ratio of 51 to 49.
- Now, performing the predictions on this dataset, we get a F1 score of 87.7%, with keeping minimum number of instances in a tree as 49, splitting subsets not smaller then

25 and limiting the maximal tree depth to 500. This score seems to be an exceptionally good score for this algorithm.

KNN Model:

- The k-nearest neighbours (KNN) technique calculates the likelihood that a data point will belong to one group, or another based on which group the data points closest to it do. KNN finds the distances between a query and each example in the data, chooses the K examples closest to the query, and then, in the case of classification, votes for the label with the highest frequency or averages the labels.
- Therefore, here as it is clear from the above explanation. We have runned the kNN model on a very balanced dataset. And distributed the model in train and test, taking nearest neighbours as 5. We get an F1 score of around 81.2%.

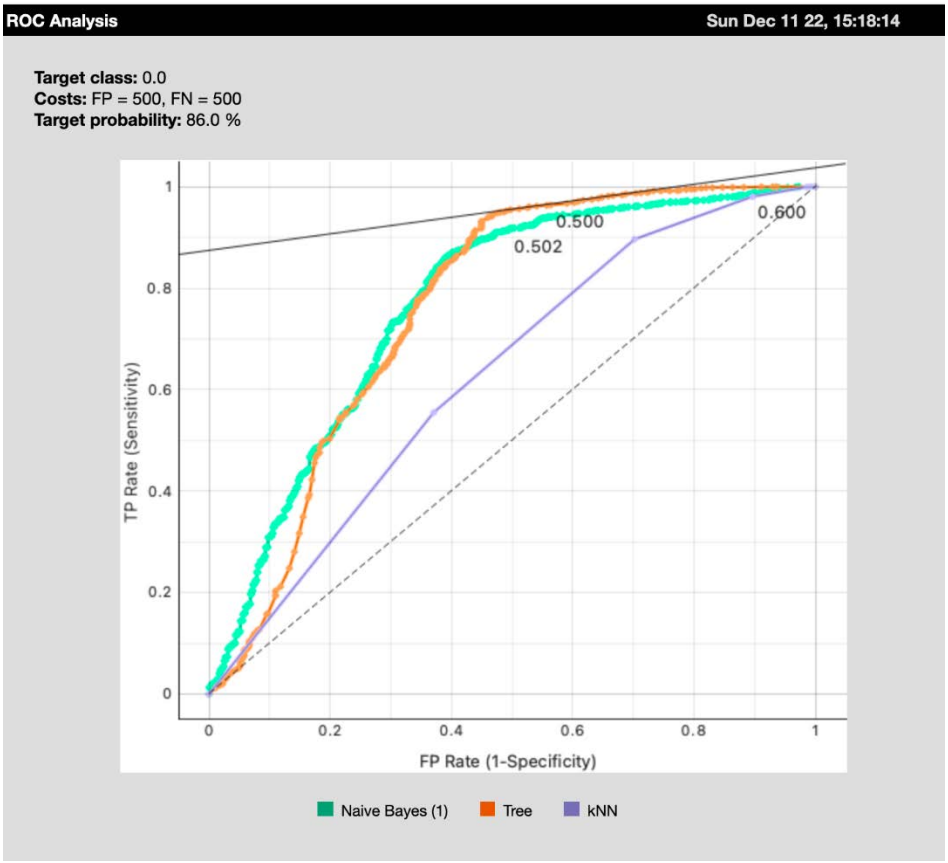


Fig : The ROC curve of Naive Bayes, Decision Tree and kNN models.

Neural Network:

A neural network is a machine learning model that is made up of interconnected nodes which are organized in layers. The connection of each neuron in each layer to the neurons in the next layers are mediated by weights, and as the model is trained, the weights are updated so as for the model to converge on the correct answers.

In this case, we deployed a multilayer perceptron, which is the most basic (least complex) form of neural network. Even though it is a basic architecture, it can achieve good performance in classification tasks like this one if it has enough neurons. That's because it can learn very complex functions that would be difficult for more simplistic machine learning methods. However, neural networks are among the least interpretable types of models, and can essentially be thought of as black boxes.

We deployed and MLP Classifier from the sklearn package. We used 4 hidden layers with 200 neurons in each layer. We used "relu" as an activation function, and 20000 as the max iterations.

Using 10-fold cross-validation to evaluate the performance of the model, we achieved an F1 score of .888.

Results:

Model summary –

MODEL	F1 Score
SVM	0.63
KNN	0.811(achieved with under sampling)
GBT	0.827
Random Forest	0.838 (achieved with oversampling)
Naïve Bayesian	0.856(achieved with under sampling)
Decision Tree	0.877 (achieved with under sampling)
Neural Networks	0.888

Recommendations:

- You can deploy our machine learning models to identify customers who are at risk of churning
- We can introduce multiple plans which best suit different customer segmentations depending on their talk time.
- We recommend introducing better international plans, because we can observe that the price difference between the people having international plan and without international plan is too less.

Discussion:

As an analyst, I would recommend the telecom company and its stakeholders to implement their pricing according to the customer segmentations and offer better international plans to their customers so that they have less churn rate. I would recommend them use our predicting models to identify the probable customer who is going to be a churn and give him some offers to make them feel that we care about the customer more than their money.

Conclusion:

After observing the data, I feel that company can improve their service a lot and it can collect different kinds of data such as the in the state feature extraction, we could have given better insights if we have information about the competitors, local tower counts in each region and context of the customer service calls. We assume that immigrant people like students and working people from other country needs international plan, so having the nationality of the customer might help us. We have not able to see the cellular data usage or any internet or data package information inside the data which could have given us some more improvements. But, with the current information we were able to generate only three recommendations.

Appendix:

Chandra Manthena: Exploratory Data Analysis, performed few statistical methods, performed Association rule mining, Naïve Bayesian and Decision tree modelling, recommendations, and report

Ashitosh Gupta: Naïve Bayesian and Decision tree, kNN and performed undersampling, recommendations

Zack Dareshori: Modelling: SVM, Random Forest, GBT, Neural Network

Nidhee Patil: data pre-processing, Principal component analysis, Logistic Regression and K-means clustering