SpeakX – Telcom Churn Prediction

Introduction

In the competitive telecom industry, keeping customers is crucial for success. Customer churn prediction helps identify which customers might stop using a company's services. By using data analytics and machine learning, businesses can study customer behavior and usage patterns to predict churn. This helps telecom companies create strategies to keep customers happy and reduce the costs of attracting new ones. Accurate churn prediction improves customer loyalty and strengthens the company's position in the market.

Libraries Used

1- Numpy: For numerical computation.

2- Pandas: For data manipulation and analysis.

3- Matplotlib: For creating static, interactive, and animated visualizations.

4- Seaborn: For easier visualization. Creating high level interface for drawing attractive and complex plots.

5- Sklearn: Scikit Learn for data mining and machine learning.

Exploratory Data Analysis

Dataset name: churn.csv

Size: 7043 rows x 21 columns

Column names: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn.

No null values

In Total Charges column there is blank values, so I have replaced with 0.0 to avoid confusion also I have changed the data type of it from object to float.

A pie chart on ‘Payment Method’.

A scatter plot which shows that Monthly charges and Total charges have linear relationship.

Then I separated the dataset into 2 parts to distinguish between Numerical and Categorical features for better visuals.

A heatmap on all the numerical features, ‘Total Charges’ have high correlation with ‘Monthly Charges’ and ‘Tenure’.

‘Tenure’ has a high correlation with ‘Total Charges’. So, ‘Total Charges’ column needs to get removed for proper analysis.

I have created a group of histograms for both numerical and categorical features differently.

Then I changed the name of the columns so that the dataset looks even better.

Now, comes the important part, in our dataset most of the columns are categorical columns and as we know that machine learning algorithms takes only numerical column so to change the categorical columns to numerical columns, I have used ‘Label Encoding’ to perform that.

At last, I have exported this dataset as CSV file and named it as ‘churn_modified.csv’ and this file is used for further analysis and machine learning model development.

Feature Engineering

So as of now I have removed the columns “Total Charges” and “Customer ID” as these will not be used for further machine learning model building and don’t contribute much for it.

And after that I checked the relevancy of the features by performing the ‘Chi-square test’ and ‘Random Forest Test’ but the outcome I got for both tests is different. So, I considered to choose all the features for the further analysis and machine learning model development.

I have done classification using 5 different machine learning algorithms, named as Logistic Regression, Random Forest Classification, Support Vector Machine, K Nearest Neighbors, Gradient Boosting Classification.

1. Logistic Regression Classification

- Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.
- Here I have imported the required libraries and modules.
- I have loaded the modified dataset. ('churn_modified.csv')
- Set X and Y values.
- Trained the model using 70 percent training data and 42 random states.
- Accuracy of the model is 81.06%
- Then I tried to predict the churning of the customer using input data and the model is successfully returning the correct answer.
- Then I have done cross validation which is used to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.
- Now Classification Report and Confusion Matrix, here TN = 1388, FP = 151, FN = 249, TP = 325, model can predict 1388 'no churn' correctly.
- Precision of the model is 68.28%
- Recall of the model is 56.62%
- The important thing in customer churn prediction is recall. This is because precision has FP (False Positive) which means that model is predicting that customer will churn but actually customer is not churning but in case of recall, it has FN (False Negative) means model is predicting that customer will not churn but actually customer is churning and it is the only important thing to find out, so here I will try to improve the recall at the cost of precision.
- Now Precision v/s Recall Curve, by observing that curve I came to know that -0.3 is the minimum threshold to improve the recall.
- And next, even in the Precision v/s Recall Graph it shows that from 0.3 the precision is sharply decreasing, and recall is sharply increasing.
- So, after adjusting new Precision is 60.26% and new Recall is 62.22%
- Then ROC curve (Receiver Operator Characteristic) using FPR and TPR which shows the performance of a binary classifier model at varying threshold values.
- At last AUC (Area under curve) means ROC and AUC score which is used to evaluate the model performance on classification. ROC AUC Score is 0.84.

2. Random Forest Classification

- Random Forest Classification combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. Decision trees.
- Here I have imported the required libraries and modules.
- I have loaded the modified dataset. ('churn_modified.csv')
- Set X and Y values.
- Trained the model using 70 percent training data and 42 random states.
- Accuracy of the model is 78.98%
- Then I tried to predict the churning of the customer using input data and the model is successfully returning the correct answer.
- Then I have done cross validation which is used to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.
- Now Classification Report and Confusion Matrix, here TN = 1399, FP = 140, FN = 304, TP = 270, model can predict 1399 'no churn' correctly.
- Precision of the model is 65.85%
- Recall of the model is 47.03%
- The important thing in customer churn prediction is recall. This is because precision has FP (False Positive) which means that model is predicting that customer will churn but actually customer is not churning but in case of recall, it has FN (False Negative) means model is predicting that customer will not churn but actually customer is churning and it is the only important thing to find out, so here I will try to improve the recall at the cost of precision.
- Now Precision v/s Recall Curve, by observing that curve I came to know that 0.9 is the minimum threshold to improve the recall.
- And next, even in the Precision v/s Recall Graph it shows that from 0.9 the precision is sharply decreasing, and recall is sharply increasing.
- So, after adjusting new Precision is 63.72% and new Recall is 48.58%
- Then ROC curve (Receiver Operator Characteristic) using FPR and TPR which shows the performance of a binary classifier model at varying threshold values.
- At last AUC (Area under curve) means ROC and AUC score which is used to evaluate the model performance on classification. ROC AUC Score is 0.69.

3. K Nearest Neighbors Classification

- K Nearest Neighbors is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
- Here I have imported the required libraries and modules.
- I have loaded the modified dataset. ('churn_modified.csv')
- Set X and Y values.
- Trained the model using 70 percent training data and 42 random states.
- Accuracy of the model is 79.27%
- Then I tried to predict the churning of the customer using input data and the model is successfully returning the correct answer.
- Then I have done cross validation which is used to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.
- Now Classification Report and Confusion Matrix, here TN = 1419, FP = 120, FN = 318, TP = 256, model can predict 1419 'no churn' correctly.
- Precision of the model is 68.08%
- Recall of the model is 44.60%
- The important thing in customer churn prediction is recall. This is because precision has FP (False Positive) which means that model is predicting that customer will churn but actually customer is not churning but in case of recall, it has FN (False Negative) means model is predicting that customer will not churn but actually customer is churning and it is the only important thing to find out, so here I will try to improve the recall at the cost of precision.
- Now Precision v/s Recall Curve, by observing that curve I came to know that 0.8 is the minimum threshold to improve the recall.
- And next, even in the Precision v/s Recall Graph it shows that from 0.8 the precision is sharply decreasing, and recall is sharply increasing.
- So, after adjusting new Precision is 65.20% and new Recall is 45.32%
- Then ROC curve (Receiver Operator Characteristic) using FPR and TPR which shows the performance of a binary classifier model at varying threshold values.
- At last AUC (Area under curve) means ROC and AUC score which is used to evaluate the model performance on classification. ROC AUC Score is 0.68.

4. Support Vector Classification

- Support Vector Classification is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.
- Here I have imported the required libraries and modules.
- I have loaded the modified dataset. ('churn_modified.csv')
- Set X and Y values.
- Trained the model using 70 percent training data and 42 random states.
- Accuracy of the model is 78.80%
- Then I tried to predict the churning of the customer using input data and the model is successfully returning the correct answer.
- Then I have done cross validation which is used to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.
- Now Classification Report and Confusion Matrix, here TN = 1439, FP = 100, FN = 348, TP = 226, model can predict 1439 'no churn' correctly.
- Precision of the model is 69.32%
- Recall of the model is 39.37%
- The important thing in customer churn prediction is recall. This is because precision has FP (False Positive) which means that model is predicting that customer will churn but actually customer is not churning but in case of recall, it has FN (False Negative) means model is predicting that customer will not churn but actually customer is churning and it is the only important thing to find out, so here I will try to improve the recall at the cost of precision.
- Now Precision v/s Recall Curve, by observing that curve I came to know that -0.7 is the minimum threshold to improve the recall.
- And next, even in the Precision v/s Recall Graph it shows that from 0.7 the precision is sharply decreasing, and recall is sharply increasing.
- So, after adjusting new Precision is 56.00% and new Recall is 60.57%
- Then ROC curve (Receiver Operator Characteristic) using FPR and TPR which shows the performance of a binary classifier model at varying threshold values.
- At last AUC (Area under curve) means ROC and AUC score which is used to evaluate the model performance on classification. ROC AUC Score is 0.79.

5. Gradient Boosting Classification

- Gradient Boosting Classification is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function.
- Here I have imported the required libraries and modules.
- I have loaded the modified dataset. ('churn_modified.csv')
- Set X and Y values.
- Trained the model using 70 percent training data and 42 random states.
- Accuracy of the model is 80.40%
- Then I tried to predict the churning of the customer using input data and the model is successfully returning the correct answer.
- Then I have done cross validation which is used to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.
- Now Classification Report and Confusion Matrix, here TN = 1399, FP = 140, FN = 274, TP = 300, model can predict 1399 'no churn' correctly.
- Precision of the model is 68.18%
- Recall of the model is 52.26%
- The important thing in customer churn prediction is recall. This is because precision has FP (False Positive) which means that model is predicting that customer will churn but actually customer is not churning but in case of recall, it has FN (False Negative) means model is predicting that customer will not churn but actually customer is churning and it is the only important thing to find out, so here I will try to improve the recall at the cost of precision.
- Now Precision v/s Recall Curve, by observing that curve I came to know that -0.3 is the minimum threshold to improve the recall.
- And next, even in the Precision v/s Recall Graph it shows that from 0.3 the precision is sharply decreasing, and recall is sharply increasing.
- So, after adjusting new Precision is 61.51% and new Recall is 60.73%
- Then ROC curve (Receiver Operator Characteristic) using FPR and TPR which shows the performance of a binary classifier model at varying threshold values.
- At last AUC (Area under curve) means ROC and AUC score which is used to evaluate the model performance on classification. ROC AUC Score is 0.84.

Comparison Table

Sr No	Model Name	Accuracy in %	Precision in %	Recall in %	Threshold	New Precision in %	New Recall in %	ROC AUC Score [0,1]
1.	Log	81.06	68.28	56.62	-0.3	60.26	62.22	0.84
2.	RF	78.98	65.85	47.03	0.9	63.72	48.58	0.69
3.	KNN	79.27	68.08	44.60	0.8	65.30	45.32	0.68
4.	SVC	78.80	69.32	39.37	-0.7	56.00	60.57	0.79
5.	GBC	80.40	68.18	52.26	-0.3	61.51	60.73	0.84

So, in conclusion by observing the comparison table: -

- RF, KNN and SVC has quite good accuracy. But Log and GBC has performed well.
- RF and KNN has not given good recall and even ROC AUC Score is less than other models.
- SVC has performed quite good and has good ROC AUC Score.
- Log and GBC has good accuracy, good recall and ROC AUC Score. Both algorithms are better than other three algorithms in customer churn prediction classification.