IFI 8420 Business Machine Learning

Final Project Report

Telecom customer churn prediction: IBM Dataset

Group 3

Shreyashi Mukhopadhyay

Business Question: Telecom Customer Churn Prediction

With the enormous increase in the number of customers using telephone services, the marketing division for a telco company wants to attract more new customers and avoid contract termination from existing customers (churn rate). For the telco company to expand its clientele, its growth rate (number of new customers) must exceed its churn rate (number of customers existing). Some of the factors that caused existing customers to leave their telco companies are better price offers, faster internet services, and a more secure online experience from other companies.

A high churn rate will adversely affect a company's profits and impede growth. This churn prediction project would be able to provide clarity to the telco company on how well it is retaining its existing customers and understand what are the underlying reasons that are causing existing customers to terminate their contract (high churn rate).

In this project, we built a model to predict how likely a customer will churn by analyzing its characteristics: (1) Demographic information (2) Account information (3) Products Information (4) Services information.

The objective is to obtain a data-driven solution that will allow us to reduce churn rates and, consequently, to increase customer satisfaction and corporation revenue.

Data Source:

The data has been sourced from Kaggle at: Telco customer churn: IBM dataset (kaggle.com)

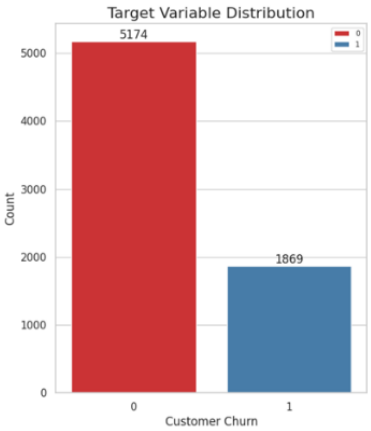
Dataset Characteristics:

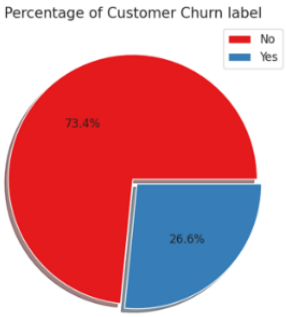
Number of rows: 7043 | Number of columns: 33

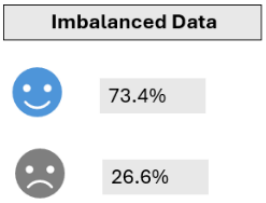
Datatypes counts: Object: 24
Int64: 6
Float64: 3

Null Values: The Churn Reason column has 5174 missing values because these customers are happy customers and hence, they have not churned.

Target Variable Distribution:



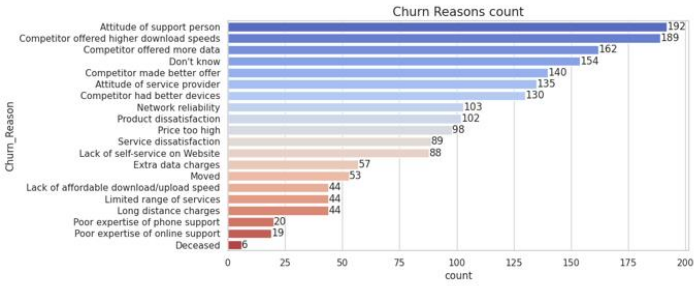




The Dataset has a positive class to negative class ratio of 5174: 1869 which is around 73.4% of positive class and 26.6% negative class. The data is imbalanced and needs to be modelled using various different modelling approaches.

EDA:

1.1 Customer churn reason

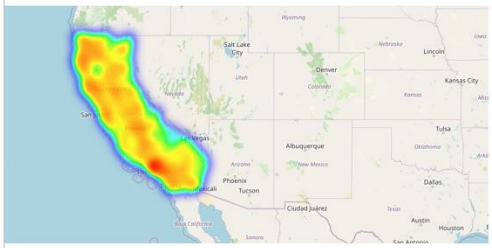


Attitude of the support person and Competitor offers were the main reasons for the customer churn.

1.2 Geographical distribution of the customers

Data Distribution

Churn values Distribution



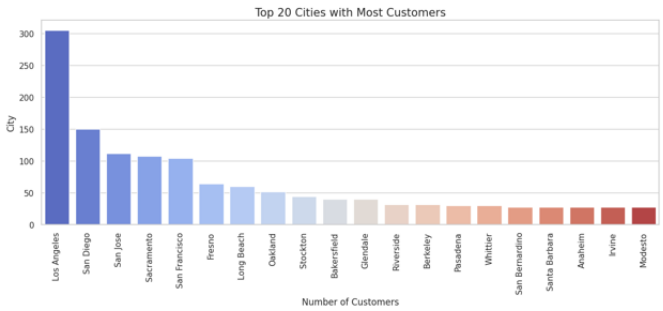
Monthly Charges Distribution

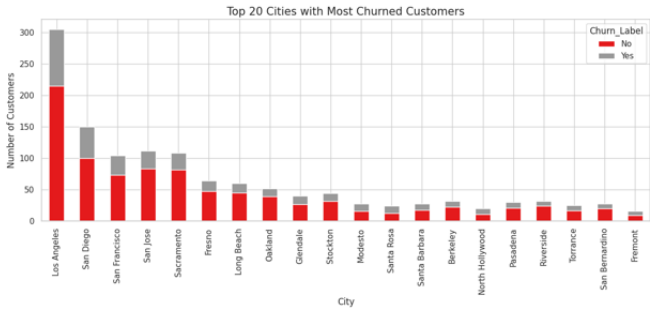
Churn Score Distribution

The customer's data is from California and Los Angeles is the city with the highest number of customers having the highest monthly charges and customer churn rate as well.

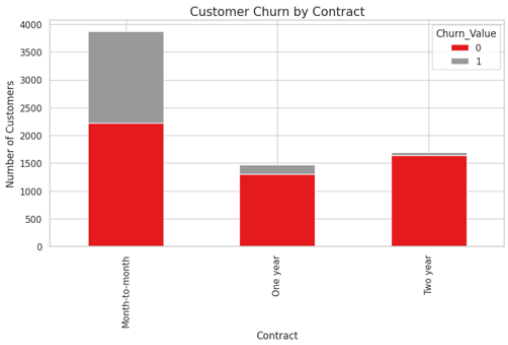
1.3 Top 20 Cities with highest customers

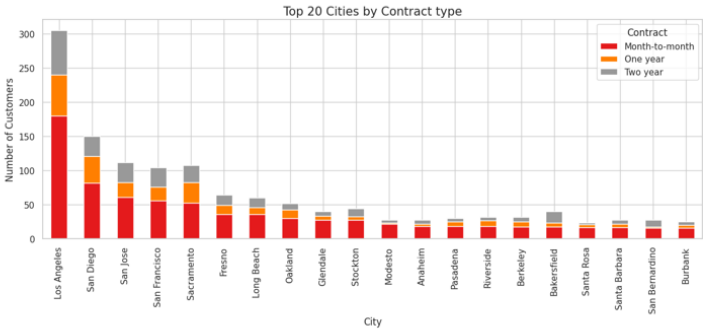
Los Angeles, San Diego, San Francisco, San Jose, Fresno, Long Beach and Oakland have the highest customers and hence the churn values are also high.





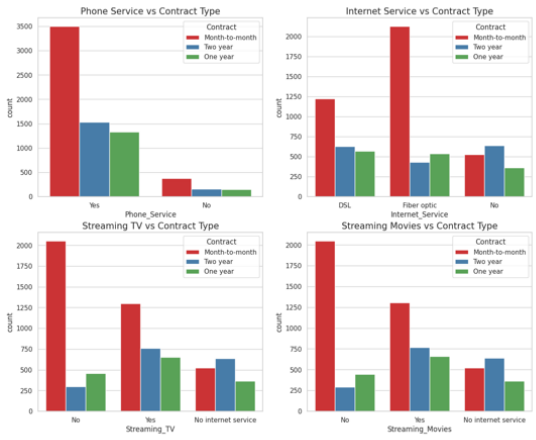
1.4 Top 20 Cities with highest customers by Contract type

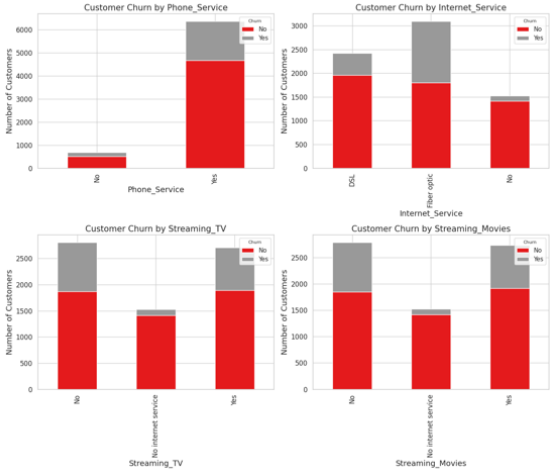




Los Angeles, San Diego, San Francisco, San Jose, Fresno, Long Beach and Oakland have the highest Monthly contracts, followed by one-year contracts and two-year contracts.

1.5 Distribution of Products by Contract type and Customer churn

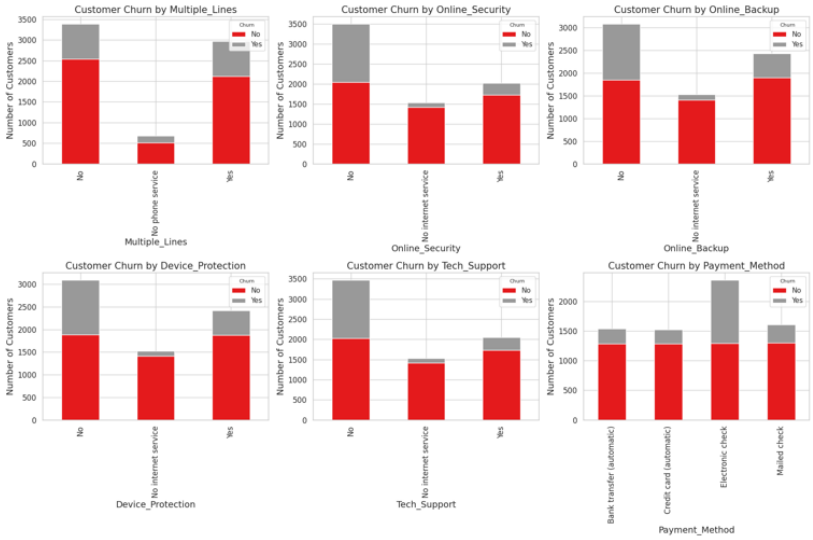




Month-to-month contracts are most popular across all categories of Phone, Internet, TV streaming and Movie Streaming Services.

Phone services, TV streaming and Movie streaming show less customer churn whereas Internet service shows high churn for Fiber optic service.

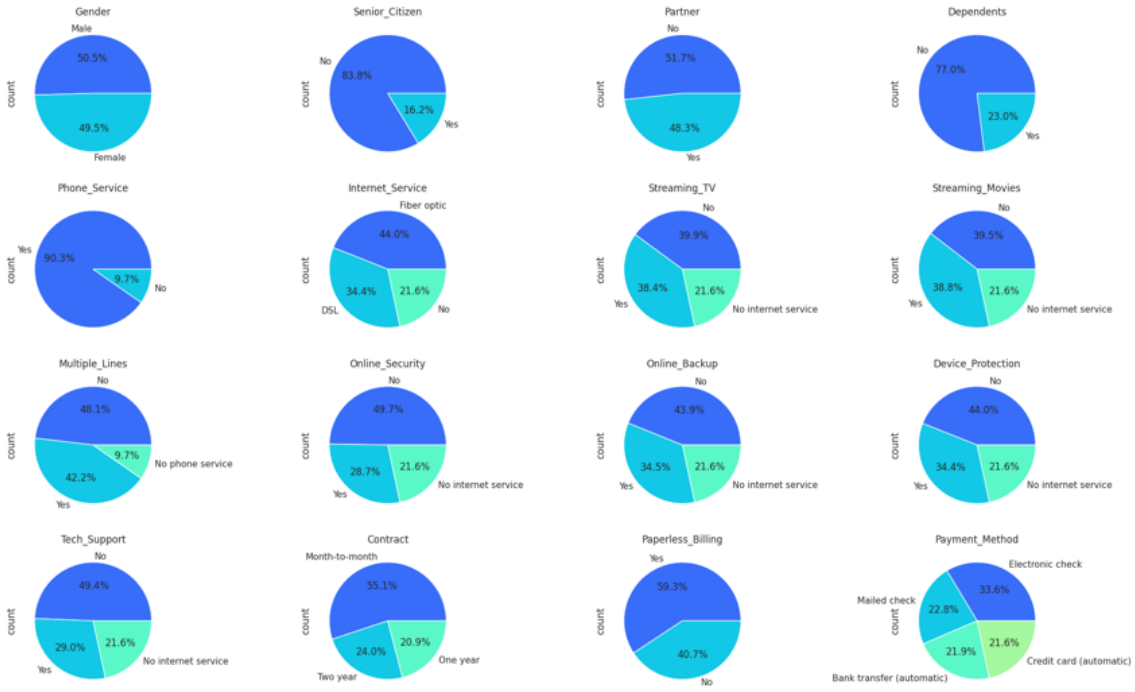
1.6 Distribution of Services by Customer Churn



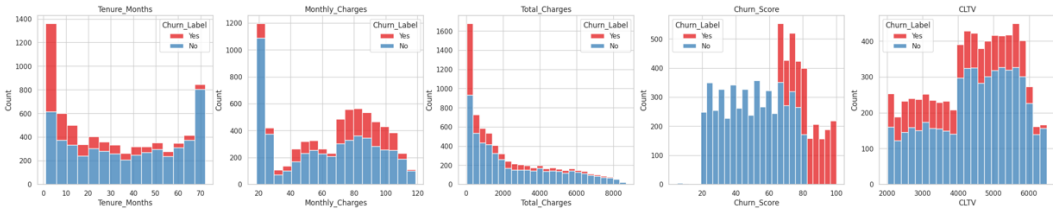
Customers prefer the Multiple lines, Online security, Online backup, Device protection, Tech support services and have very low churn value. Only the Electronic checks payment method has a high churn value.

1.7 Distribution of Customer attributes, Products and services

Equal Distribution of Gender, Few Senior Citizens, Low Dependents, High Month to month contracts, High subscription for Phone Service, High subscription for Multiple lines, Higher Fiber optic customers, High paperless billing preference.

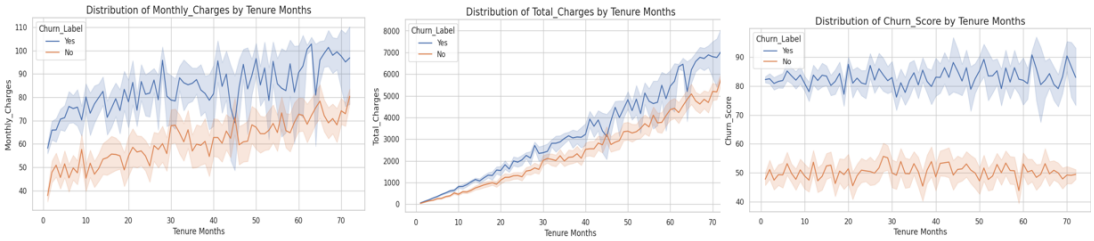


1.8 Distribution of Continuous Numerical Variables



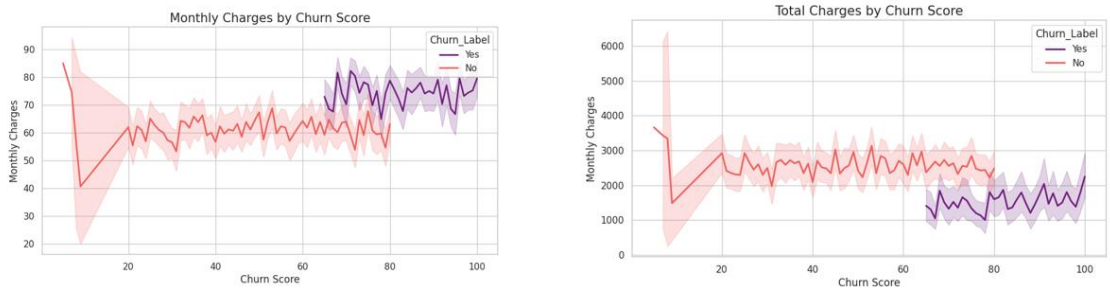
- Customer churn is high for less than 1 year contract, and it decreases after that.
- Customer churn increases for higher monthly charges.
- Customer churn is high for lesser total charges and decreases after that.
- Churn score is higher for a score 65 and above.
- Churn values are equal across all the Customer lifetime values.

1.9 Customer Churn Continuous Attributes and Trends



- Monthly charges increases with the tenure months.
- Total charges increases with increase in tenure months.
- Churn score trends are constant and around 45-55 for no-churn customers whereas it is around 80-100 for customers who churn.

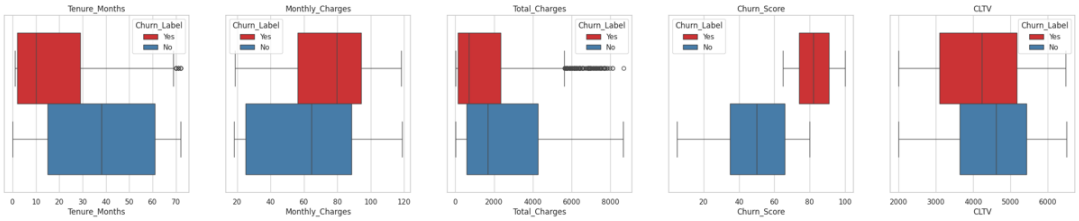
1.10 Monthly Charges and Total Charges by Churn Score



Customers churn score increases beyond Monthly charges between $60-70.

Customer churn score is higher when Total charges between $1000 - 2000.

1.11 Box Plots for Outlier detection:



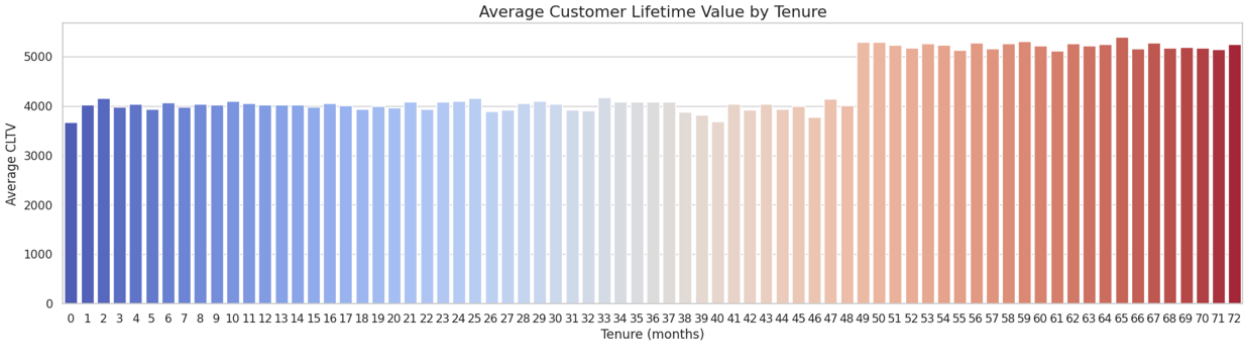
There are very few outliers for the Tenure months and the Total charges features.

We will keep these data points to utilize the information for customer behavior from these data points.

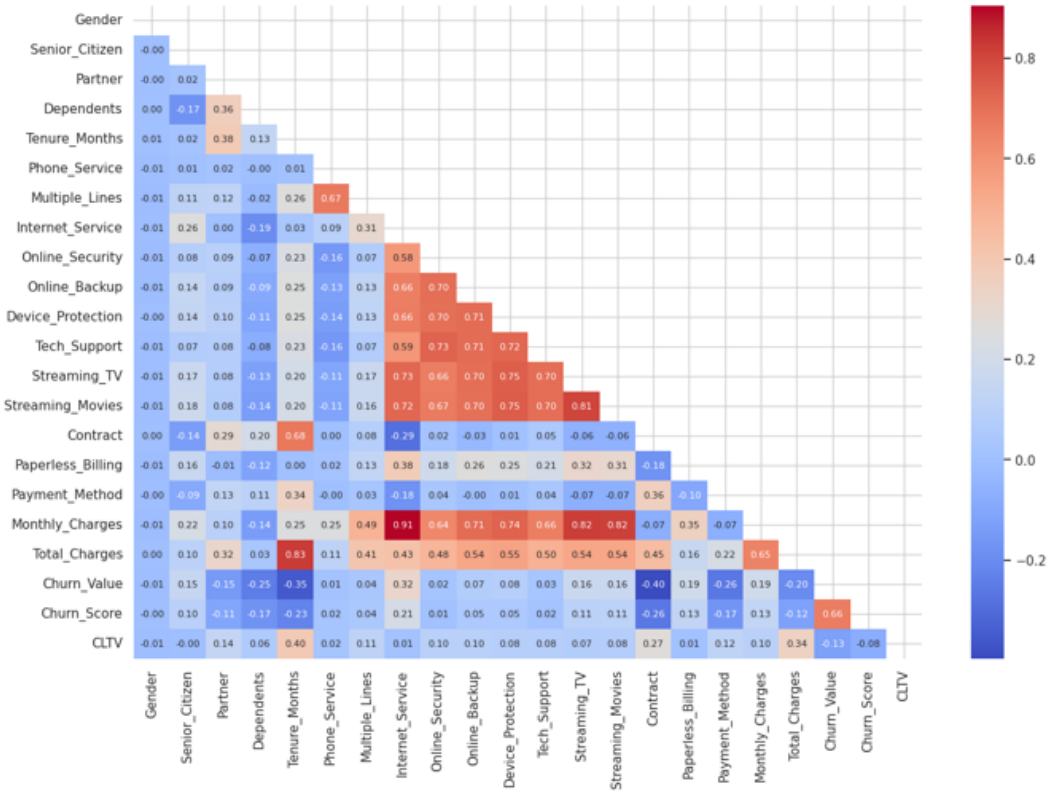
1.12 Distribution of Contract types for the Continuous numeric attributes

The month-to-month contract type has the highest value across all continuous customer attributes across all contract types as the customer's preferred product followed by Two-year plans and One-year plans.

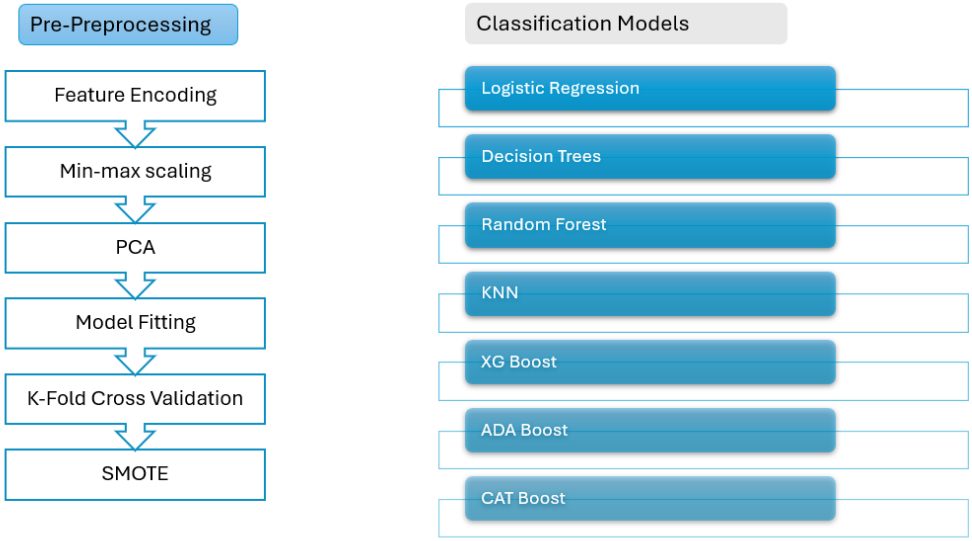
1.13 Average Customer Lifetime Value with Tenure months: The Customer Lifetime value is constant around 4000 for the first 48 months and then it suddenly increases to around 5000 beyond the 48 months.



2.1 Correlation Heatmap: There is a high correlation between most of the products and services. We can observe that a customer who buys one product definitely buys the associated services and leading up to a high correlation between the features. Many of the products and services show high correlation which shows the customer's tendency to bundle products and services.



3.0 Machine Learning: Pre-processing and Model Fitting



3.1 Feature Encoding: All the categorical features have been manually encoded to mimic a ordinal encoding process.

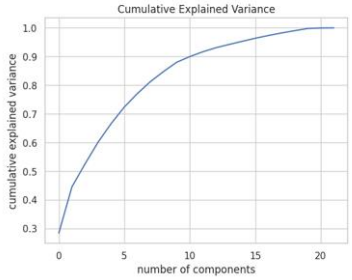
3.2 Min-max Scaling: Since all the continuous numerical variables are not normally distributed, the min-max scaling has been applied to scale the numeric features in the same range.

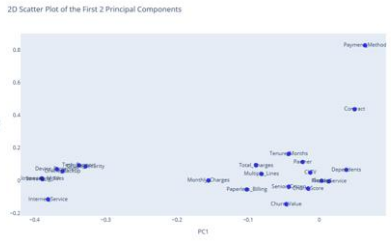
3.3 PCA for Dimensionality Reduction: PCA has been applied for dimensionality reduction to choose the best features. A total of 21 components have been identified that explain the total variability of the model. Please see the PCA plots below:

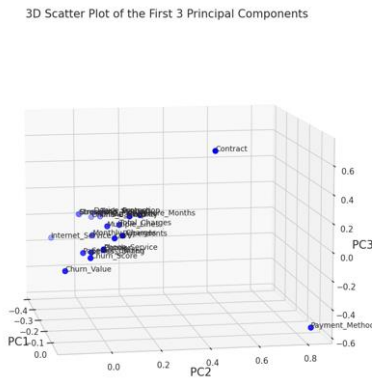
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22
Standard deviation	0.9402	0.7093	0.5002	0.4868	0.4507	0.4235	0.3836	0.3611	0.3308	0.3173	...	0.2098	0.1881	0.1844	0.1828	0.1735	0.1662	0.1580	0.1525	0.0769	0.0357
Proportion of variance	0.2834	0.1613	0.0802	0.0760	0.0651	0.0575	0.0472	0.0418	0.0351	0.0323	...	0.0141	0.0113	0.0109	0.0107	0.0096	0.0089	0.0080	0.0075	0.0019	0.0004
Cumulative proportion	0.2834	0.4448	0.5250	0.6010	0.6661	0.7236	0.7708	0.8126	0.8477	0.8800	...	0.9308	0.9421	0.9530	0.9637	0.9734	0.9822	0.9902	0.9977	0.9996	1.0000

3 rows × 22 columns

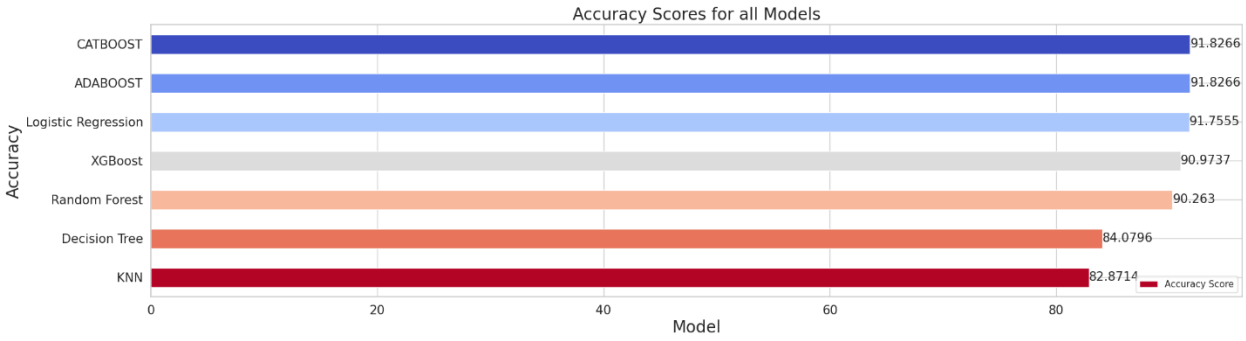
n_Components = 21



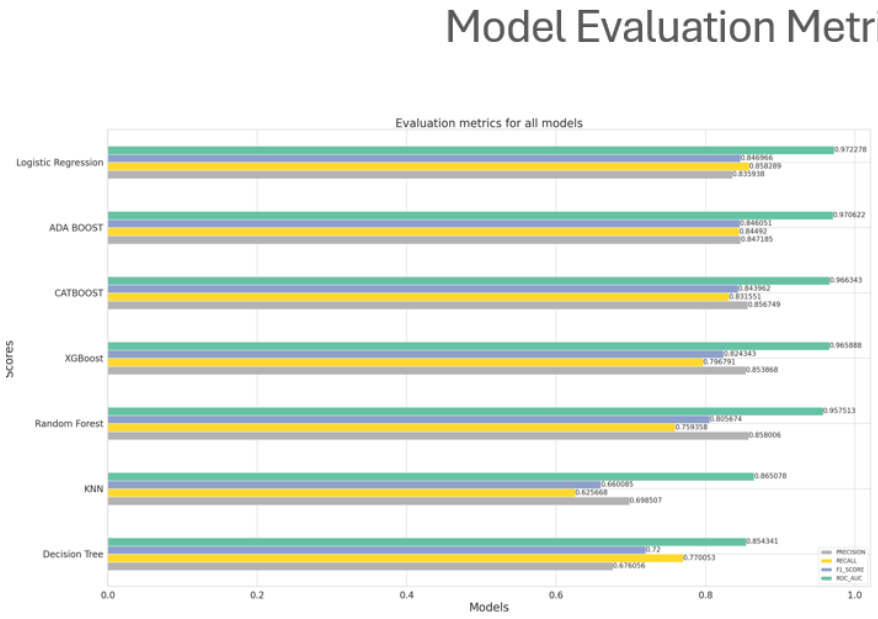




3.4 Model fitting with PCA: Model performance: CATBoost and ADABoost models have the highest accuracies followed by Logistic Regression.



3.5 Model performance with evaluation metrics: Best model is Logistic regression and model selection criteria is Recall and ROC AUC Score.



MODEL	PRECISION	RECALL	F1_SCORE	ROC_AUC
Decision Tree	0.676056	0.770053	0.720000	0.854341
KNN	0.698507	0.625668	0.660085	0.865078
Random Forest	0.858006	0.759358	0.805674	0.957513
XGBoost	0.853868	0.796791	0.824343	0.965888
CATBOOST	0.856749	0.831551	0.843962	0.966343
ADA BOOST	0.847185	0.844920	0.846051	0.970622
Logistic Regression	0.835938	0.858289	0.846966	0.972278

We can observe that Logistic Regression has the highest Precision, Recall, F1 Score and ROC AUC Score among all the other models. The model selection criteria are Recall and ROC AUC Score.

3.6 Model Testing for Logistic Regression and ADA Boost: Both models are not able to predict the classes accurately, so we need to conduct further model experiments to arrive at a better model.

Model testing on sample data

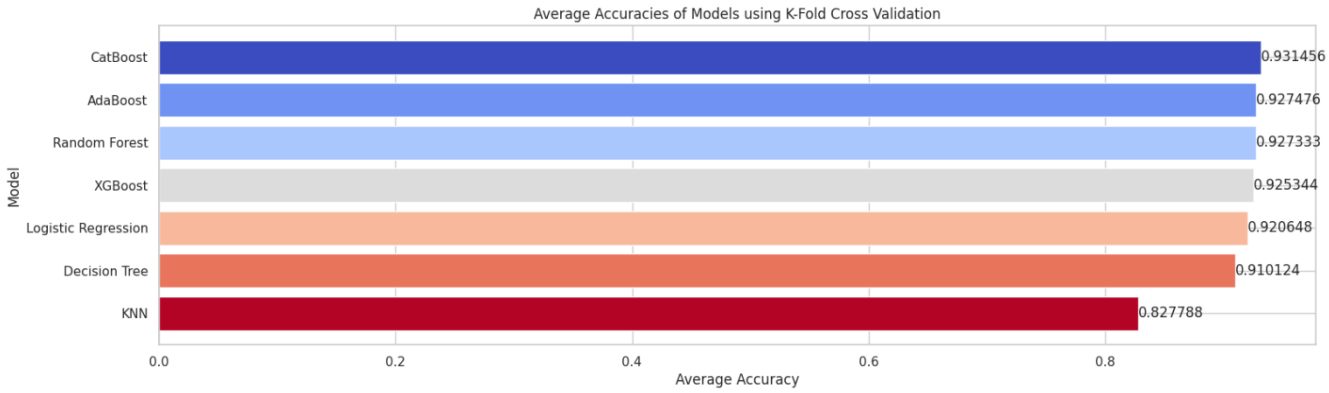
Logistic Regression

```
[241] # Test a sample data on logistic regression model

      # Create a sample data point
      sample_data = X.iloc[10]

      # Make a prediction on the sample data
      prediction = model_lr.predict([sample_data])

      # Print the prediction
      print(f"Prediction for the sample data: {prediction}")

      # Compare with the orginal value of the churn value from y

      actual_value = y.iloc[10]
      print(f"Actual label of the churn value: {actual_value}")

      Prediction for the sample data: [0]
      Actual label of the churn value: 1
```

ADA Boost

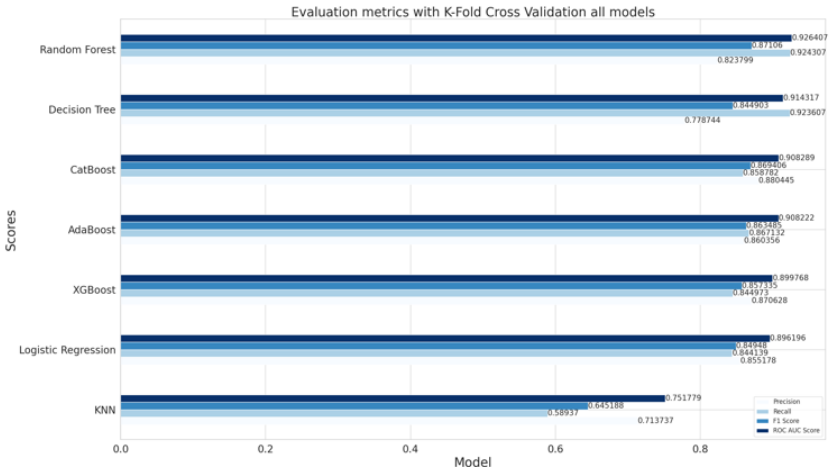
```
# Test the model for ADA Boost model

# Test a sample data on ADA Boost model

# Create a sample data point
sample_data = X.iloc[10]

# Make a prediction on the sample data
prediction = model_ada.predict([sample_data])

# Print the prediction
print(f"Prediction for the sample data: {prediction}")

# Compare with the orginal value of the churn value from y

actual_value = y.iloc[10]
print(f"Actual label of the churn value: {actual_value}")

Prediction for the sample data: [0]
Actual label of the churn value: 1
```

4.0 K-fold Cross validation Model performance accuracy: We can observe that the boosting trees CATBoost and ADA Boost have the best accuracies followed by random forest.



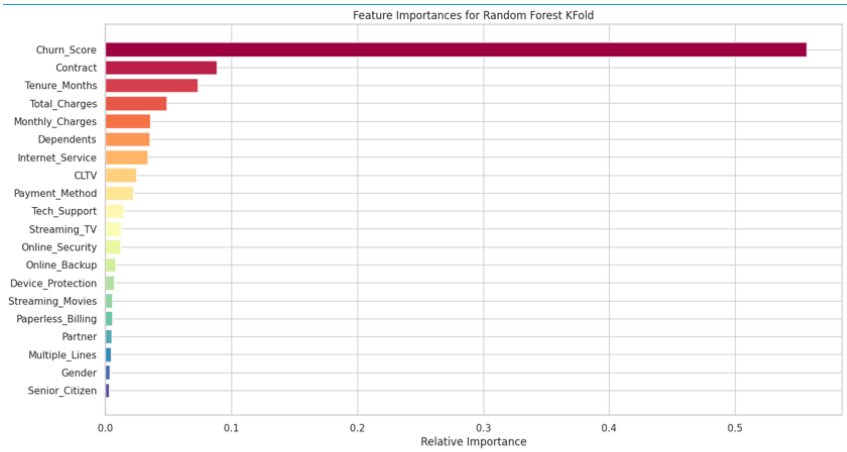
4.1 K-fold cross validation evaluation metrics: Best model is Random Forest since Random forest performs the best with the highest precision, recall, F1 score and Roc-Auc score.

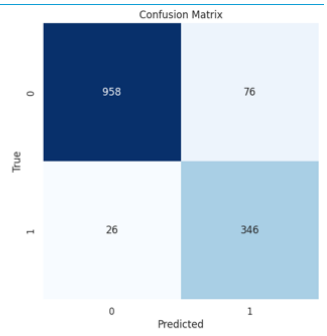
Evaluation metrics K-fold cross validation

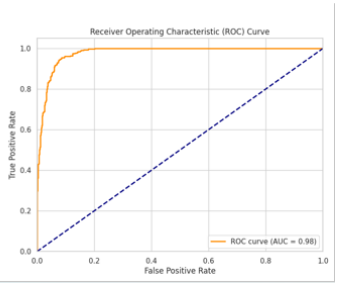


Model	Precision	Recall	F1 Score	ROC AUC Score
Random Forest	0.823799	0.924307	0.871060	0.926407
Decision Tree	0.778744	0.923607	0.844903	0.914317
CatBoost	0.880445	0.858782	0.869406	0.908289
AdaBoost	0.860356	0.867132	0.863485	0.908222
XGBoost	0.870628	0.844973	0.857335	0.899768
Logistic Regression	0.855178	0.844139	0.849480	0.896196
KNN	0.713737	0.589370	0.645188	0.751779

4.2 Feature Importances for Random Forest with K-Fold Cross Validation:








```
Average Accuracy: 0.9273328541199712
Average Precision: 0.8237988209809695
Average Recall: 0.9243067611447424
Average F1 Score: 0.8710603165192733
Average ROC AUC Score: 0.9264066049150225
```

4.3 K-fold Random Forest model testing with sample data: The model is able to accurately classify the sample data into the appropriate class.

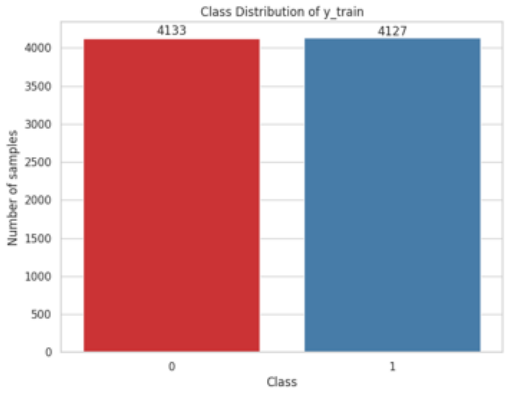
Model testing for Random Forest K-fold on sample data

```
# Test the random forest model on a sample data

# Create a sample data point
sample_data = X.iloc[10]

# Make a prediction on the sample data
prediction = model_rf_kfold.predict([sample_data])

# Print the prediction
print(f"Prediction for the sample data: {prediction}")

# Compare with the orginal value of the churn value from y

actual_value = y.iloc[10]
print(f"Actual label of the churn value: {actual_value}")
```

```
Prediction for the sample data: [1]
Actual label of the churn value: 1
```

5.0 SMOTE: Since the target class is imbalanced, SMOTE has been applied to the data to balance both the classes and test the models on this balanced data.

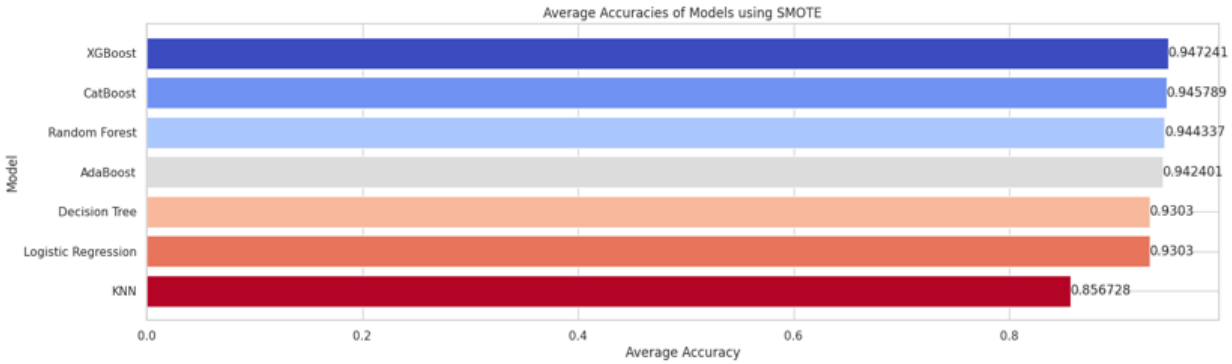
Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does not change the number of majority cases.



The new instances are not just copies of existing minority cases. Instead, the algorithm takes samples of the feature space for each target class and its nearest neighbors. The algorithm then generates new examples that combine features of the target case with features of its neighbors.

This approach increases the features available to each class and makes the samples more general.

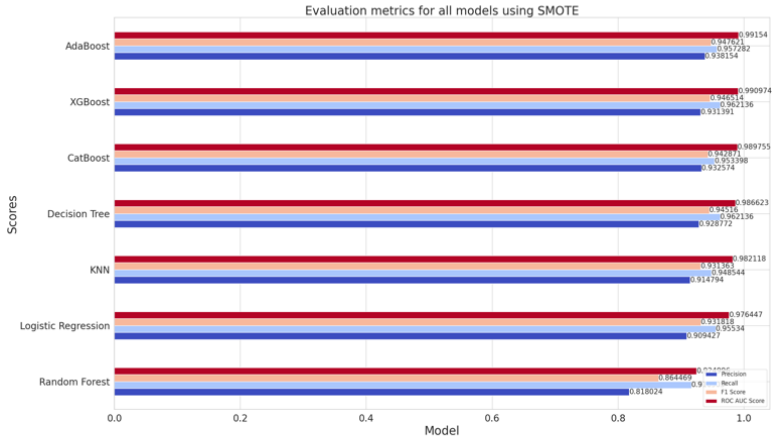
5.1 Model Accuracies and comparison with SMOTE:



XGBoost model performed with highest accuracy followed by CatBoost

5.2 Model Evaluation comparison and model selection:

Model selection criteria Recall & ROC AUC: Best model ADABoost



Model	Precision	Recall	F1 Score	ROC AUC Score
Random Forest	0.818024	0.916505	0.864469	0.924996
Logistic Regression	0.909427	0.955340	0.931818	0.976447
KNN	0.914794	0.948544	0.931363	0.982118
Decision Tree	0.928772	0.962136	0.945160	0.986623
CatBoost	0.932574	0.953398	0.942871	0.989755
XGBoost	0.931391	0.962136	0.946514	0.990974
AdaBoost	0.938154	0.957282	0.947621	0.991540

ADA Boost Model performs the best with highest Precision, Recall, F1 Score and ROC-AUC Score

5.3 Model testing on sample data: The ADABoost model is able to classify the sample data in the appropriate class.

Model Testing for Sample data with ADA Boost SMOTE

- Model predicts the classes accurately.

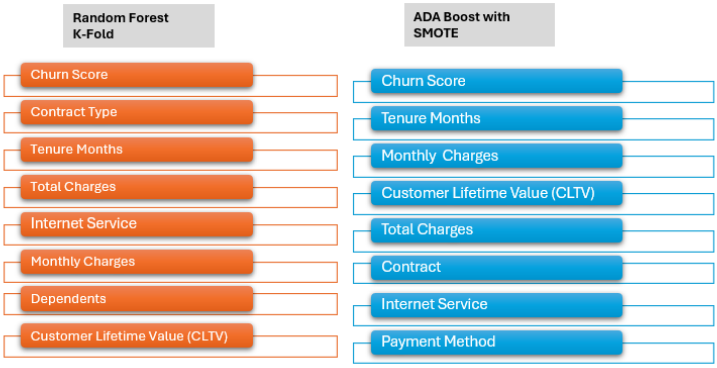
```
[178] # Test the AdaBoost model on a sample data point from SMOTE

      # Create a sample data point from SMOTE
      sample_data_smote = X_resampled.iloc[10]

      # Make a prediction on the sample data
      prediction_smote = model_ada_smote.predict([sample_data_smote])

      # Print the prediction
      print(f"Prediction for the sample data: {prediction_smote}")

      # Compare with the orginal value of the churn value from y_resampled

      actual_value_smote = y_resampled.iloc[10]
      print(f"Actual label of the churn value: {actual_value_smote}")


      Prediction for the sample data: [1]
      Actual label of the churn value: 1
```

6.0 Conclusion: The best models for our data are K-fold Random Forest and ADA Boost with SMOTE.

K Fold Random Forest
VS
ADA Boost Smote

Top features for Churn Prediction

Tree based Algorithms and Boosting Algorithms performed best.



The feature importances of both the model suggests that the most important features for our data are:

- Churn Score
- Contract type
- Tenure Months
- Total Charges
- Monthly Charges
- Internet service
- Dependents
- Customer Lifetime Value (CLTV)