Project Report

On

Telecom Churn Prediction

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics from **C-DAC Hyderabad**

Guided by:

Mr. Ganga Prasad

Presented by:

Yogesh Bakle	**PRN: 220950325006**
Amit Awad	**PRN: 220950325005**
Abhishek Mookherjee	**PRN: 220950325001**
Amardeep	**PRN: 220950325004**
Bishwajit Sen	**PRN: 220950325007**

Centre of Development of Advanced Computing

(C-DAC), Hyderabad

CERTIFICATE

This is to certify that,

Yogesh Bakle
Amit Awad
Abhishek Mookherjee
Amardeep
Bishwajit Sen

Have successfully completed their project on

Telecom Churn Prediction

Under the guidance of Mr. Ganga Prasad

Project Guide **Project Supervisor**

Course Coordinator:

Mr. Sadhu Sreenivas.

ACKNOWLEDGEMENT

This project **"Telecom Churn Prediction"** was a great learning experience for us and we are submitting this work to AdvancedComputing Training School (CDAC Hyderabad).

We all are very glad to mention the name of **Mr. Ganga Prasad** for his valuableguidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to Mr. Sadhu Sreenivas (Course Coordinator, Hyderabad) for her guidance and support whenever necessary while doing this course Post Graduate Diploma in **Big Data Analytics (PG -DBDA)**
Through C-DAC Hyderabad.

Our most heartfelt thanks goes to **Mrs.Sadhu Sreenivas** (Course Coordinator, **PG - DBDA**) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC ACTS, Hyderabad.

From:
Yogesh Bakle (220950325006)
Amit Awad (220950325005)
Abhishek Mookherjee (220950325001)
Amardeep (220950325004)
Bishwajit Sen (220950325007)

TABLE OF CONTENTS

1. Abstract

Telecom churn prediction is the process of identifying customer who are likely to determine their subscription or switch to a different service provider. This is an important task for telecom companies as it helps them retain customer and reduce revenue loss.

Various techniques, such as machine Learning algorithms and statistical models, are used to predict customer churn. These techniques analyze various factors, such as customer demographics, usage pattern, and customer service interactions, to determine the likelihood of churn. The accuracy of churn prediction models can be improved by continuously updating the data and refining the models based on customer behavior and feedback.

Overall, telecom churn prediction is an essential tool for telecom companies to manage customer retention and Maximize revenue.

2. Introduction and Overview of Project

Telecom churn prediction project is a data-driven approach that helps telecom companies to identify customers who are at risk of churning or leaving the service. The project involves analyzing customer behavior, usage patterns, and demographics data to develop predictive models that can anticipate churn. The goal of this project is to help telecom companies understand why customers leave, which factors contribute to churn, and how to retain customers.

The project typically involves several steps, starting with data collection and preprocessing, followed by feature engineering and model development. The data is usually collected from multiple sources, such as call logs, customer service interactions, billing data, and customer surveys. The data is then preprocessed to remove any noise, missing values, and outliers.

Feature engineering is the process of selecting and transforming the most relevant features that are likely to influence customer churn. Features may include customer demographics, usage patterns, and behavior, such as the number of calls made, the duration of calls, the types of plans used, and the frequency of complaints.

The next step involves building predictive models using machine learning algorithms, such as logistic regression, decision trees, and random forests. These models are trained on historical data to predict customer churn based on the selected features. The accuracy of the models is evaluated using performance metrics such as accuracy, precision, recall, and F1 score.

Finally, the project concludes with model deployment and ongoing monitoring of customer behavior. The predictive models are integrated into the telecom company's system to provide real-time predictions of customer churn. The model's performance is continuously monitored and updated based on customer feedback and new data to ensure that the model remains accurate and effective.

Overall, telecom churn prediction project is an essential tool for telecom companies to manage customer retention, reduce churn, and maximize revenue. The project's success depends on the quality of data, the accuracy of predictive models, and the effectiveness of retention strategies implemented by the telecom company.

3. Problem Statement

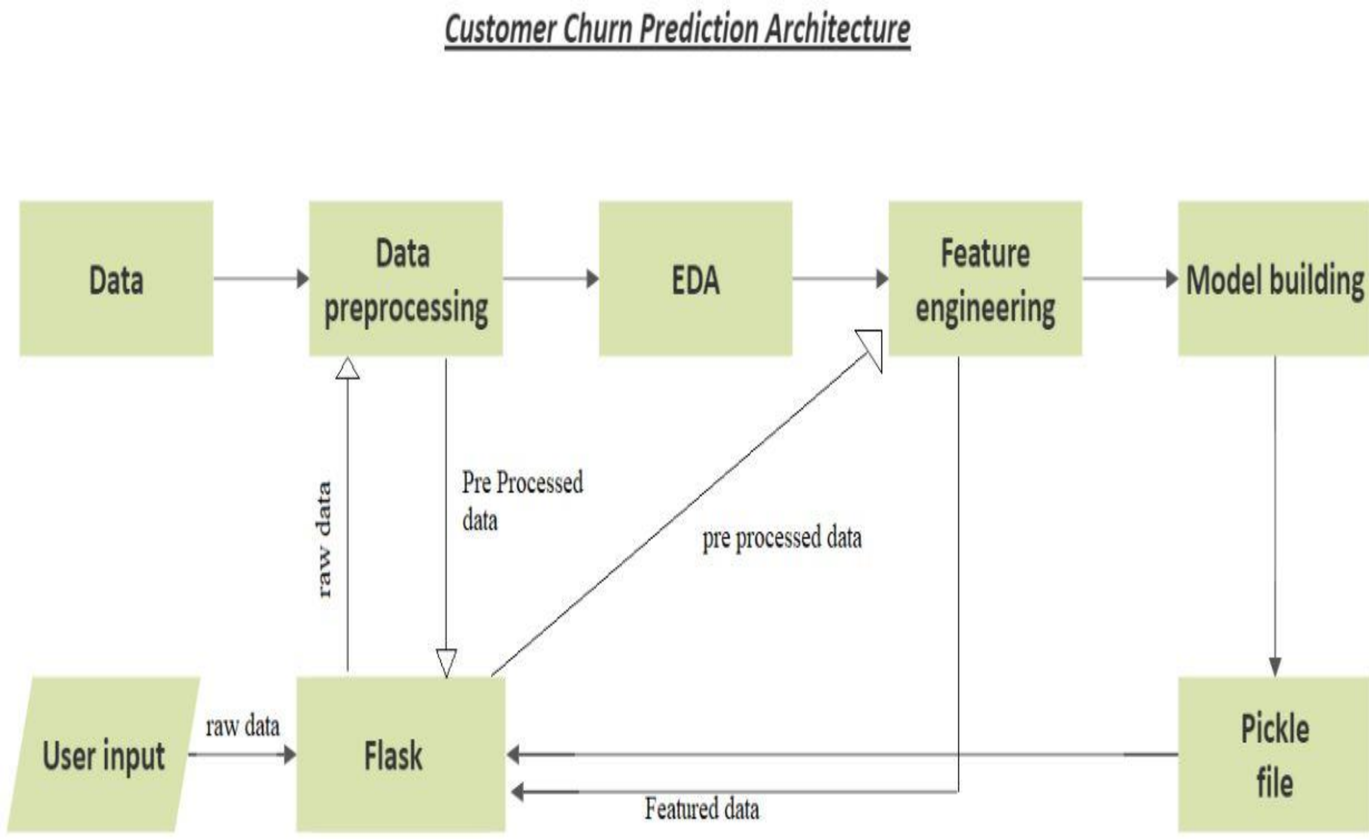
Customer churn means shifting one service provider to its competitor. Customer churn is one of the biggest fears of any industry particularly for telecom industry. Although there are many reasons for customer churn, some of the measure reasons are service dissatisfaction, costly subscription and better alternatives. Hence predicting churn in the telecom industry is very important.

Objectives of the Project:

The primary objectives of telecom churn prediction is to reduce customer churn and improve customer retention rate for telecom companies.

1. Identify factors that are associated with higher churn rates.
2. Develop predictive model to identify customers at high risk of churn
3. Improve customer retention rates.

Work Flow of Project



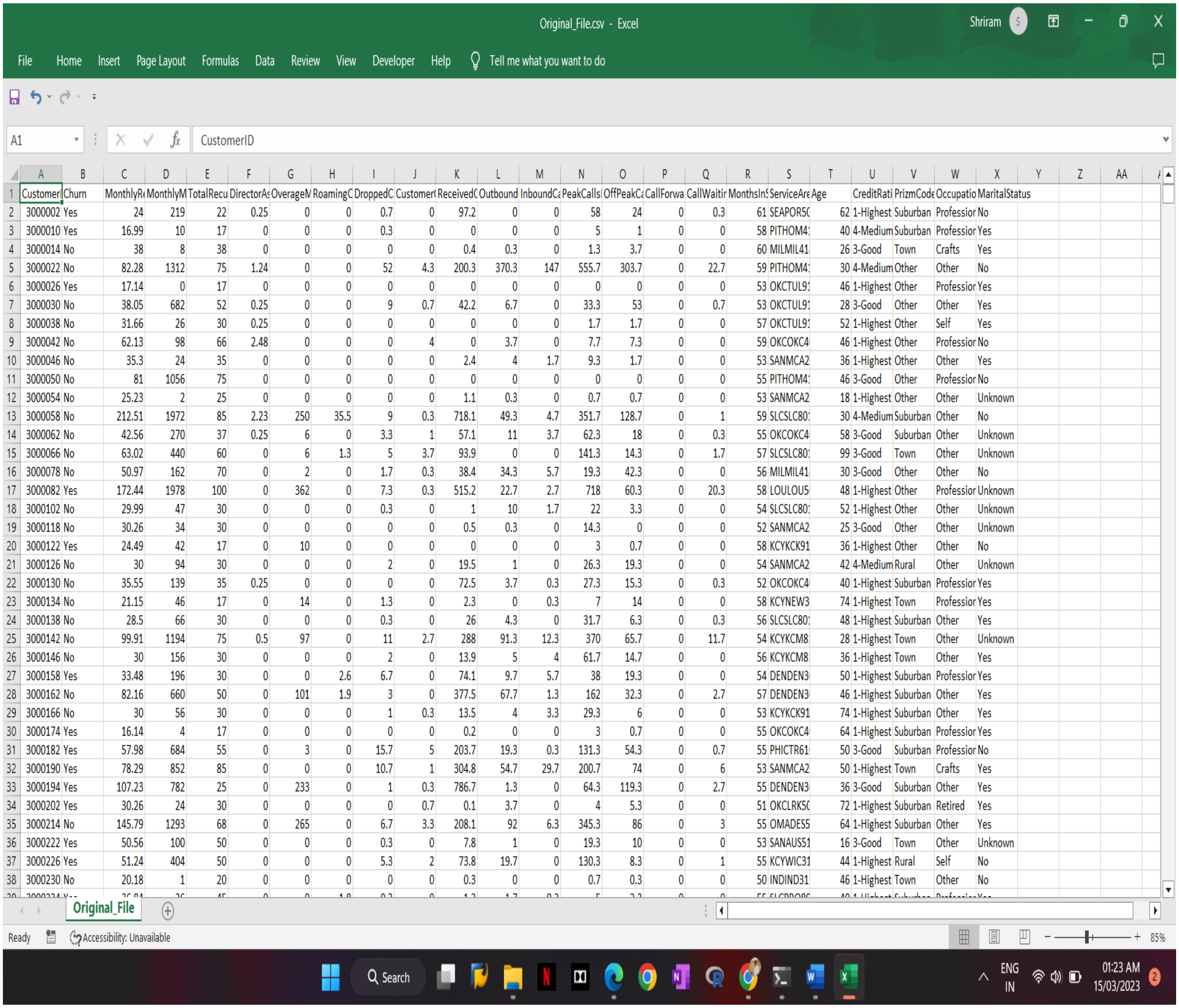
4. Dataset Description

The data set used in this article is available in the **Kaggle** and contains **Twenty-four columns (independent variables)** that indicate the **characteristics of the clients** of a fictional telecommunications corporation. The Churn column (**response variable**) indicates whether the customer departed within the last month or not. The class No includes the clients that did not leave the company last month, while the class Yes contains the clients that decided to terminate their relations with the company. The objective of the analysis is to obtain **the relation between the customer's characteristics and the churn**.

- Dataset Structure: 51047 observations (rows), 24 features (variables)
- Data Type: Three datatypes in this dataset: objects, integers, and float
- Imbalanced dataset: 36336 (71.18% of cases) employees did not leave the organization while 14711 (28.81% of cases) did leave the organization making us dataset to be considered imbalanced since more people stay in the organization than they actually leave

Name	Description
Churn	Customer leaving the Service (0=no, 1=yes)
Monthly Revenue	Numerical Value
Monthly Minutes	Numerical Value
Total Recurring Charge	Numerical Value
Overage Minutes	Numerical Value
Roaming Calls	Numerical Value
Dropped Calls	Numerical Value

Customer Care Calls	Numerical Value
Received Calls	Numerical Value
Outbound Calls	Numerical Value
Inbound Calls	Numerical Value
Peak Calls In Out	Numerical Value
Off Peak Calls In Out	Numerical Value
Call Forwarding Calls	Numerical Value
Call Waiting Calls	Numerical Value
Months In Service	Numerical Value
Service Area	Object Type
Age	Numerical Value
Credit Rating	Object Type
Prizm Code	Object Type
Occupation	Object Type
Marital Status	Object Type



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	CustomerI	Churn	MonthlyRe	MonthlyM	TotalRecu	DirectorAs	OverageM	RoamingC	DroppedC	Customer	ReceivedC	Outbound	InboundCa	PeakCalls	OffPeakCa	CallForwa	CallWaitir	MonthsIn	ServiceAre	Age	CreditRati	PrizmCode	Occupatio	MaritalStatus
2	3000002	Yes	24	219	22	0.25	0	0	0.7	0	97.2	0	0	58	24	0	0.3	61	SEAPOR5C	62	1-Highest	Suburban	Professior	No
3	3000010	Yes	16.99	10	17	0	0	0	0.3	0	0	0	0	5	1	0	0	58	PITHOM41	40	4-Medium	Suburban	Professior	Yes
4	3000014	No	38	8	38	0	0	0	0	0	0.4	0.3	0	1.3	3.7	0	0	60	MILMIL41	26	3-Good	Town	Crafts	Yes
5	3000022	No	82.28	1312	75	1.24	0	0	52	4.3	200.3	370.3	147	555.7	303.7	0	22.7	59	PITHOM41	30	4-Medium	Other	Other	No
6	3000026	Yes	17.14	0	17	0	0	0	0	0	0	0	0	0	0	0	0	53	OKCTUL91	46	1-Highest	Other	Professior	Yes
7	3000030	No	38.05	682	52	0.25	0	0	9	0.7	42.2	6.7	0	33.3	53	0	0.7	53	OKCTUL91	28	3-Good	Other	Other	Yes
8	3000038	No	31.66	26	30	0.25	0	0	0	0	0	0	0	1.7	1.7	0	0	57	OKCTUL91	52	1-Highest	Other	Self	Yes
9	3000042	No	62.13	98	66	2.48	0	0	0	4	0	3.7	0	7.7	7.3	0	0	59	OKCOKC4	46	1-Highest	Other	Professior	No
10	3000046	No	35.3	24	35	0	0	0	0	0	2.4	4	1.7	9.3	1.7	0	0	53	SANMCA2	36	1-Highest	Other	Other	Yes
11	3000050	No	81	1056	75	0	0	0	0	0	0	0	0	0	0	0	0	55	PITHOM41	46	3-Good	Other	Professior	No
12	3000054	No	25.23	2	25	0	0	0	0	0	1.1	0.3	0	0.7	0.7	0	0	53	SANMCA2	18	1-Highest	Other	Other	Unknown
13	3000058	No	212.51	1972	85	2.23	250	35.5	9	0.3	718.1	49.3	4.7	351.7	128.7	0	1	59	SLCSLC80	30	4-Medium	Suburban	Other	No
14	3000062	No	42.56	270	37	0.25	6	0	3.3	1	57.1	11	3.7	62.3	18	0	0.3	55	OKCOKC4	58	3-Good	Suburban	Other	Unknown
15	3000066	No	63.02	440	60	0	6	1.3	5	3.7	93.9	0	0	141.3	14.3	0	1.7	57	SLCSLC80	99	3-Good	Town	Other	Unknown
16	3000078	No	50.97	162	70	0	2	0	1.7	0.3	38.4	34.3	5.7	19.3	42.3	0	0	56	MILMIL41	30	3-Good	Other	Other	No
17	3000082	Yes	172.44	1978	100	0	362	0	7.3	0.3	515.2	22.7	2.7	718	60.3	0	20.3	58	LOULOU5	48	1-Highest	Other	Professior	Unknown
18	3000102	No	29.99	47	30	0	0	0	0.3	0	1	10	1.7	22	3.3	0	0	54	SLCSLC80	52	1-Highest	Other	Other	Unknown
19	3000118	No	30.26	34	30	0	0	0	0	0	0.5	0.3	0	14.3	0	0	0	52	SANMCA2	25	3-Good	Other	Other	Unknown
20	3000122	Yes	24.49	42	17	0	10	0	0	0	0	0	0	3	0.7	0	0	58	KCYKCK91	36	1-Highest	Other	Other	No
21	3000126	No	30	94	30	0	0	0	2	0	19.5	1	0	26.3	19.3	0	0	54	SANMCA2	42	4-Medium	Rural	Other	Unknown
22	3000130	No	35.55	139	35	0.25	0	0	0	0	72.5	3.7	0.3	27.3	15.3	0	0.3	52	OKCOKC4	40	1-Highest	Suburban	Professior	Yes
23	3000134	No	21.15	46	17	0	14	0	1.3	0	2.3	0	0.3	7	14	0	0	58	KCYNEW3	74	1-Highest	Town	Professior	Yes
24	3000138	No	28.5	66	30	0	0	0	0.3	0	26	4.3	0	31.7	6.3	0	0.3	56	SLCSLC80	48	1-Highest	Suburban	Other	Yes
25	3000142	No	99.91	1194	75	0.5	97	0	11	2.7	288	91.3	12.3	370	65.7	0	11.7	54	KCYKCM8	28	1-Highest	Town	Other	Unknown
26	3000146	No	30	156	30	0	0	0	2	0	13.9	5	4	61.7	14.7	0	0	56	KCYKCM8	36	1-Highest	Town	Other	Yes
27	3000158	Yes	33.48	196	30	0	0	2.6	6.7	0	74.1	9.7	5.7	38	19.3	0	0	54	DENDEN3	50	1-Highest	Suburban	Professior	Yes
28	3000162	No	82.16	660	50	0	101	1.9	3	0	377.5	67.7	1.3	162	32.3	0	2.7	57	DENDEN3	46	1-Highest	Suburban	Other	Yes
29	3000166	No	30	56	30	0	0	0	1	0.3	13.5	4	3.3	29.3	6	0	0	53	KCYKCK91	74	1-Highest	Suburban	Other	Yes
30	3000174	Yes	16.14	4	17	0	0	0	0	0	0.2	0	0	3	0.7	0	0	55	OKCOKC4	64	1-Highest	Suburban	Professior	Yes
31	3000182	Yes	57.98	684	55	0	3	0	15.7	5	203.7	19.3	0.3	131.3	54.3	0	0.7	55	PHICTR61	50	3-Good	Suburban	Professior	No
32	3000190	Yes	78.29	852	85	0	0	0	10.7	1	304.8	54.7	29.7	200.7	74	0	6	53	SANMCA2	50	1-Highest	Town	Crafts	Yes
33	3000194	Yes	107.23	782	25	0	233	0	1	0.3	786.7	1.3	0	64.3	119.3	0	2.7	55	DENDEN3	36	3-Good	Suburban	Other	Yes
34	3000202	Yes	30.26	24	30	0	0	0	0	0.7	0.1	3.7	0	4	5.3	0	0	51	OKCLRK5C	72	1-Highest	Suburban	Retired	Yes
35	3000214	No	145.79	1293	68	0	265	0	6.7	3.3	208.1	92	6.3	345.3	86	0	3	55	OMADES5	64	1-Highest	Suburban	Other	Yes
36	3000222	Yes	50.56	100	50	0	0	0	0.3	0	7.8	1	0	19.3	10	0	0	53	SANAUS51	16	3-Good	Town	Other	Unknown
37	3000226	Yes	51.24	404	50	0	0	0	5.3	2	73.8	19.7	0	130.3	8.3	0	1	55	KCYWIC31	44	1-Highest	Rural	Self	No
38	3000230	No	20.18	1	20	0	0	0	0	0	0.3	0	0	0.7	0.3	0	0	50	INDIND31	46	1-Highest	Town	Other	No

Fig. Dataset

4.1 Label Encoding:

In machine learning, we usually deal with datasets that contain multiple labels in one or more than one columns. These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labelled in words.

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.
We have done label encoding by
using get_dummies function of pandas library. In our dataset, we applied label encoding on following categorical columns:

1. Service Area
2. Credit Rating
3. Prizm Code
4. Occupation
5. Marital Status
6. Churn

5. Data Pre-processing and Cleaning

DATA CLEANING:

- Missing Data: There are 7 (MonthlyRevenue, MonthlyMinutes, TotalRecuringCharge, OverageMinutes, RoamingCalls, ServiceArea, Age) Columns in which missing values are present.

- Data Type: We have two data types in this dataset: Categorical and Numerical.

- The label "Churn" is the label in our dataset and we would like to find out why customer are leaving the service!

- Label Encoding: To make the data understandable or in human readable form, the training data is often labeled in words. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine- readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated.

- In our dataset there are 24 variables however, sum features just have one data level that do not make sense for our research such as CustomerID, DirectedAssistedCalls and ServiceArea number doesn't have meaning in analyzing result so we are deleted these features.

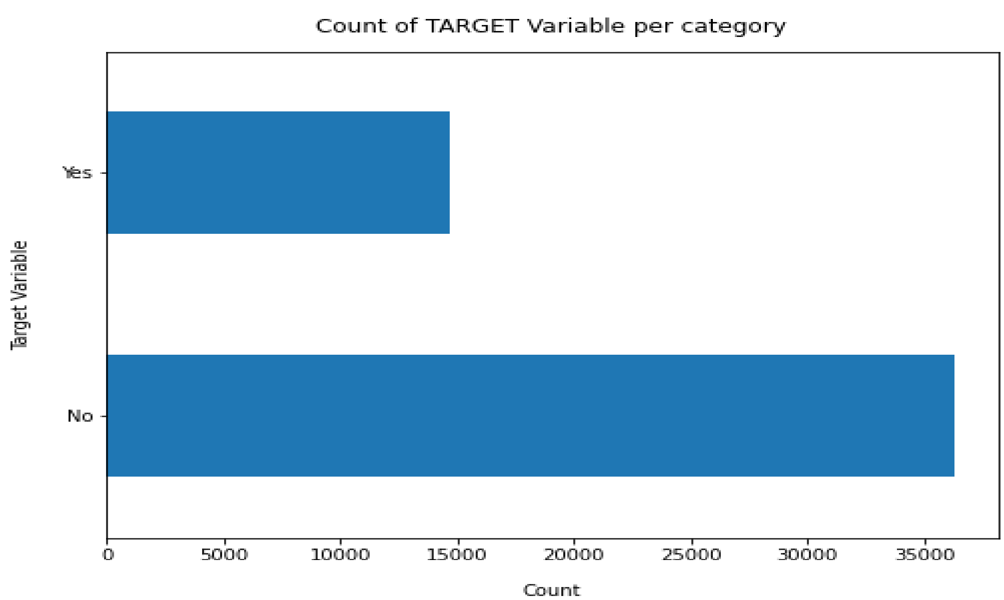
- Imbalanced dataset: 36336 (71.18% of cases) employees did not leave the service while 14711 (28.81% of cases) did leave the service making our dataset to be considered imbalanced since more peoplestay in the service than they actually leave.

6. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

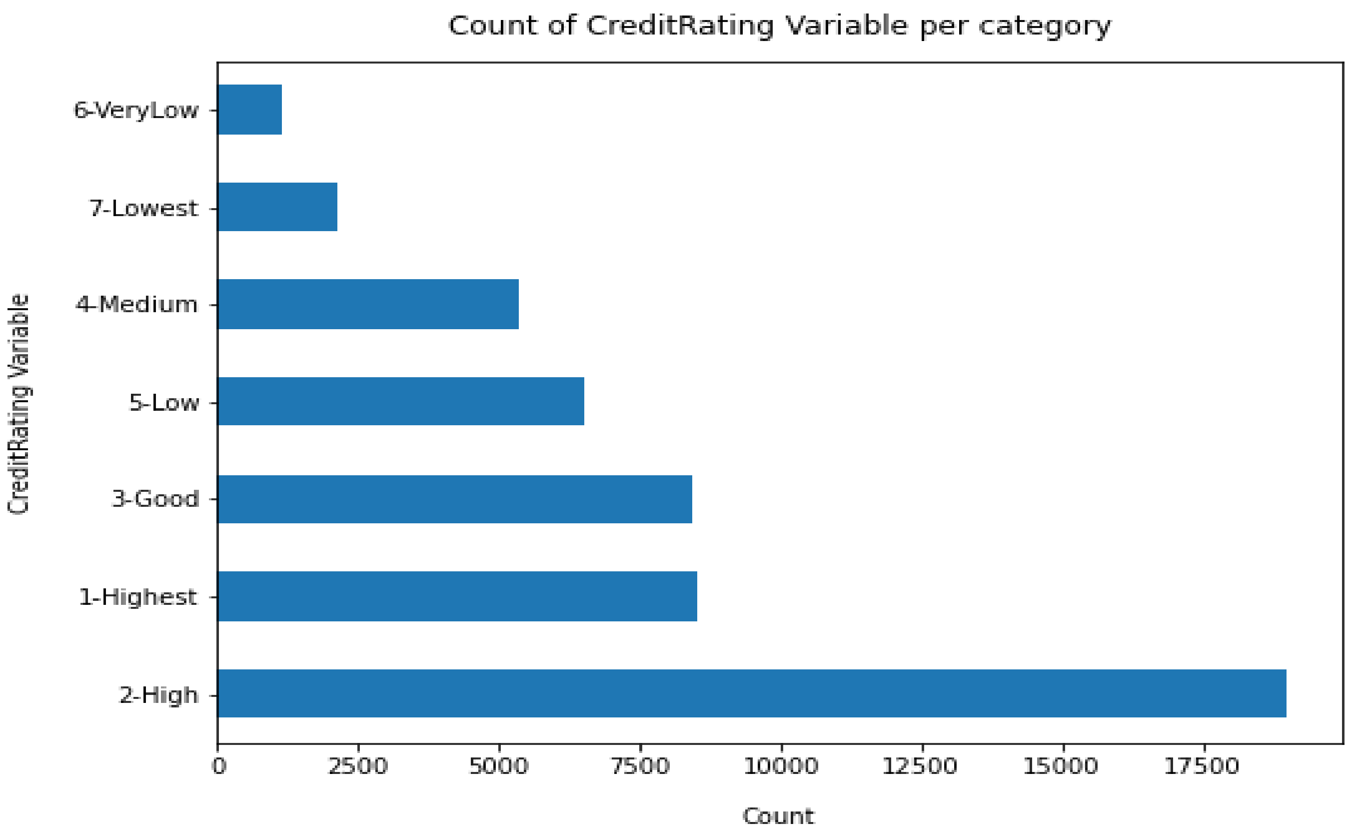
Univariate Data Analysis:

In our dataset target variable is Churn. Our target variable is binary so that it isnecessary to check data is balanced or not.



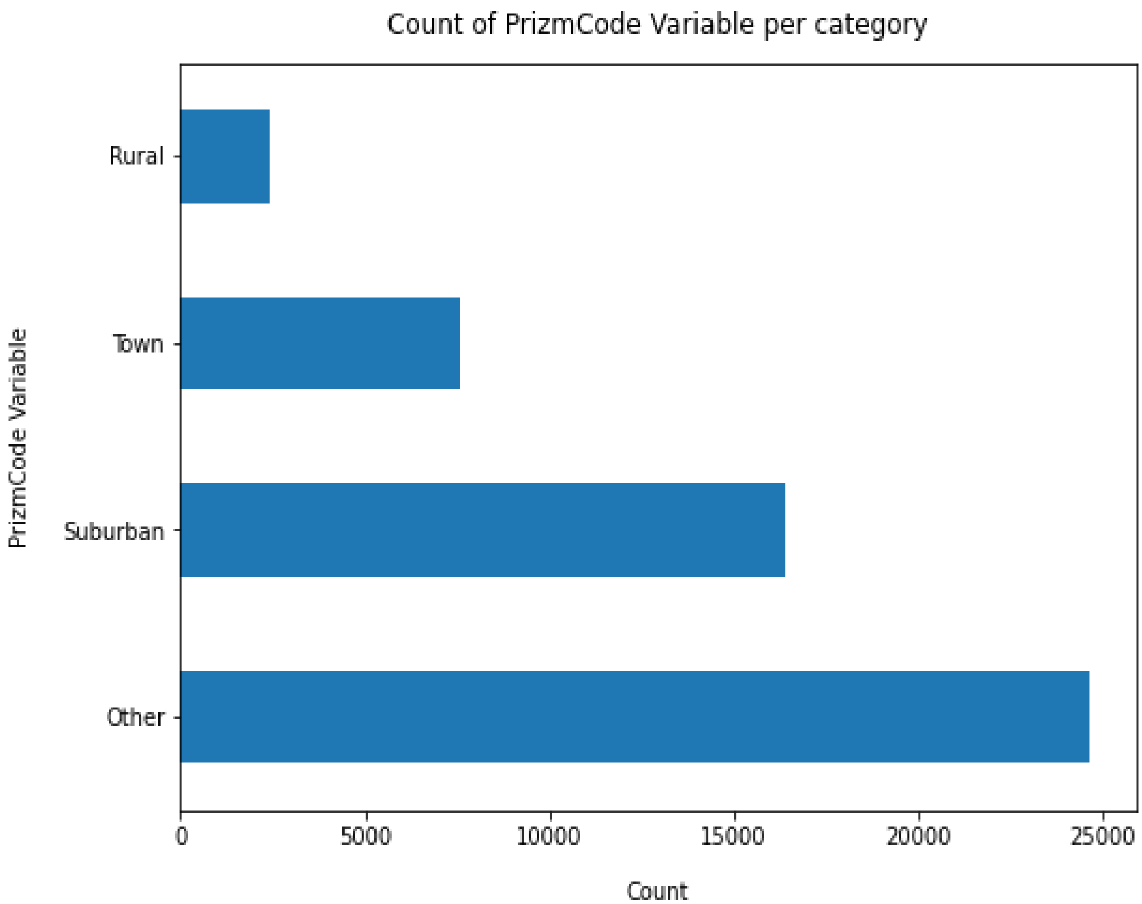
Conclusion: From above bar chart, we come to know that the % of No churn is more than % of Yes churn.

- **Analyze the Credit Rating Column.**



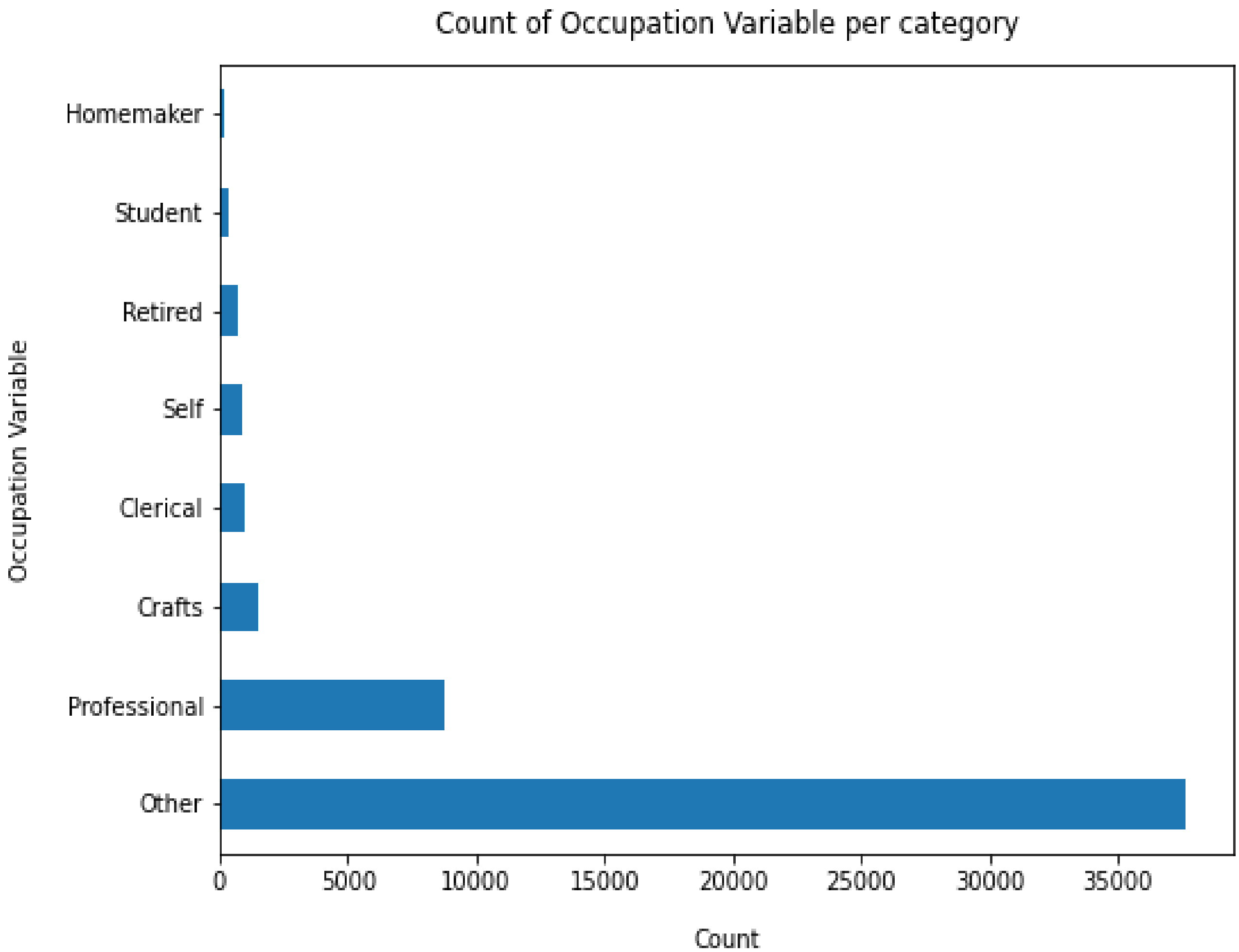
Conclusion: From the above chart we can say that maximum number of customer have high credit rating.

- **Analyze the Prizm Code Column.**



Conclusion: Most of the customers falls under the other category.

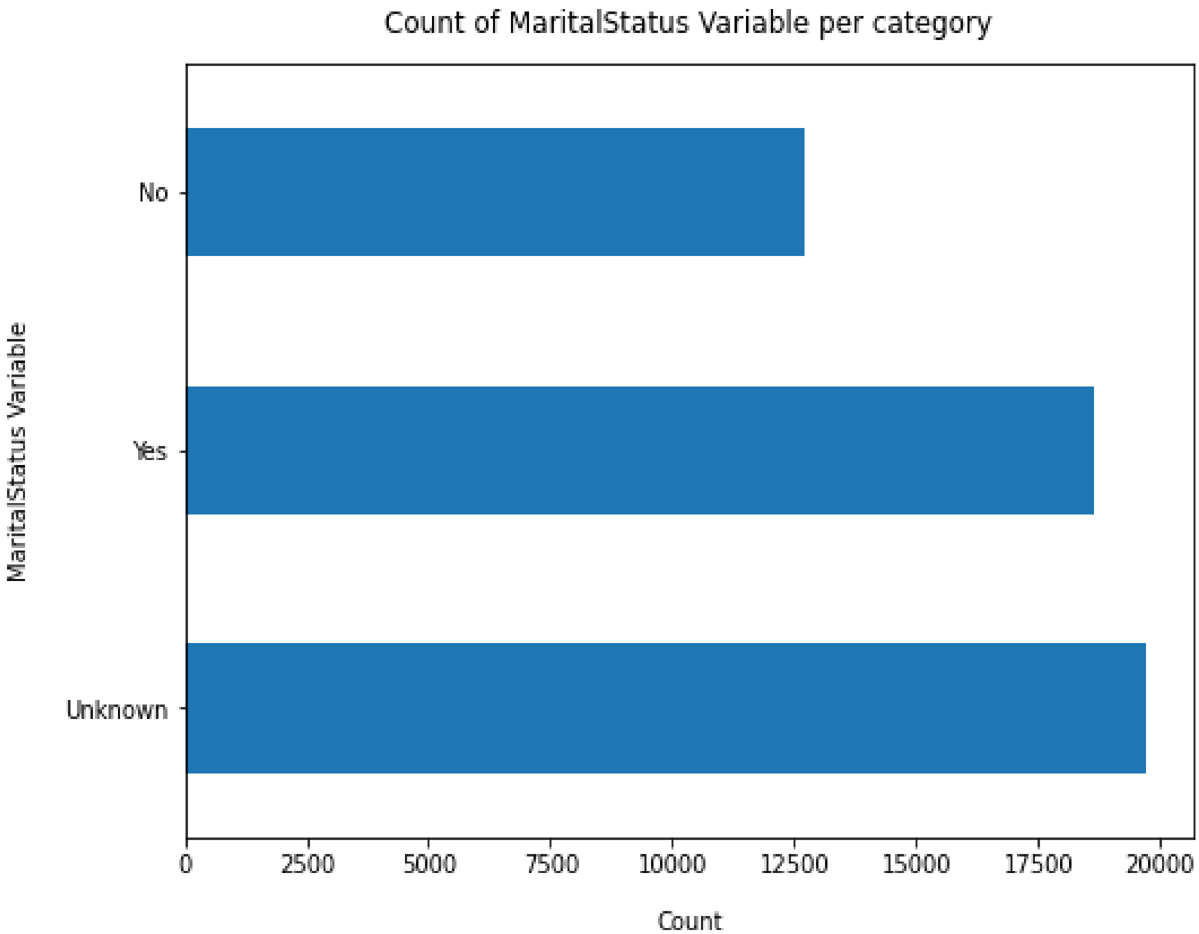
- **Analyze Occupation Column.**



Conclusion: Most of the customers Occupation falls under the other category.

.

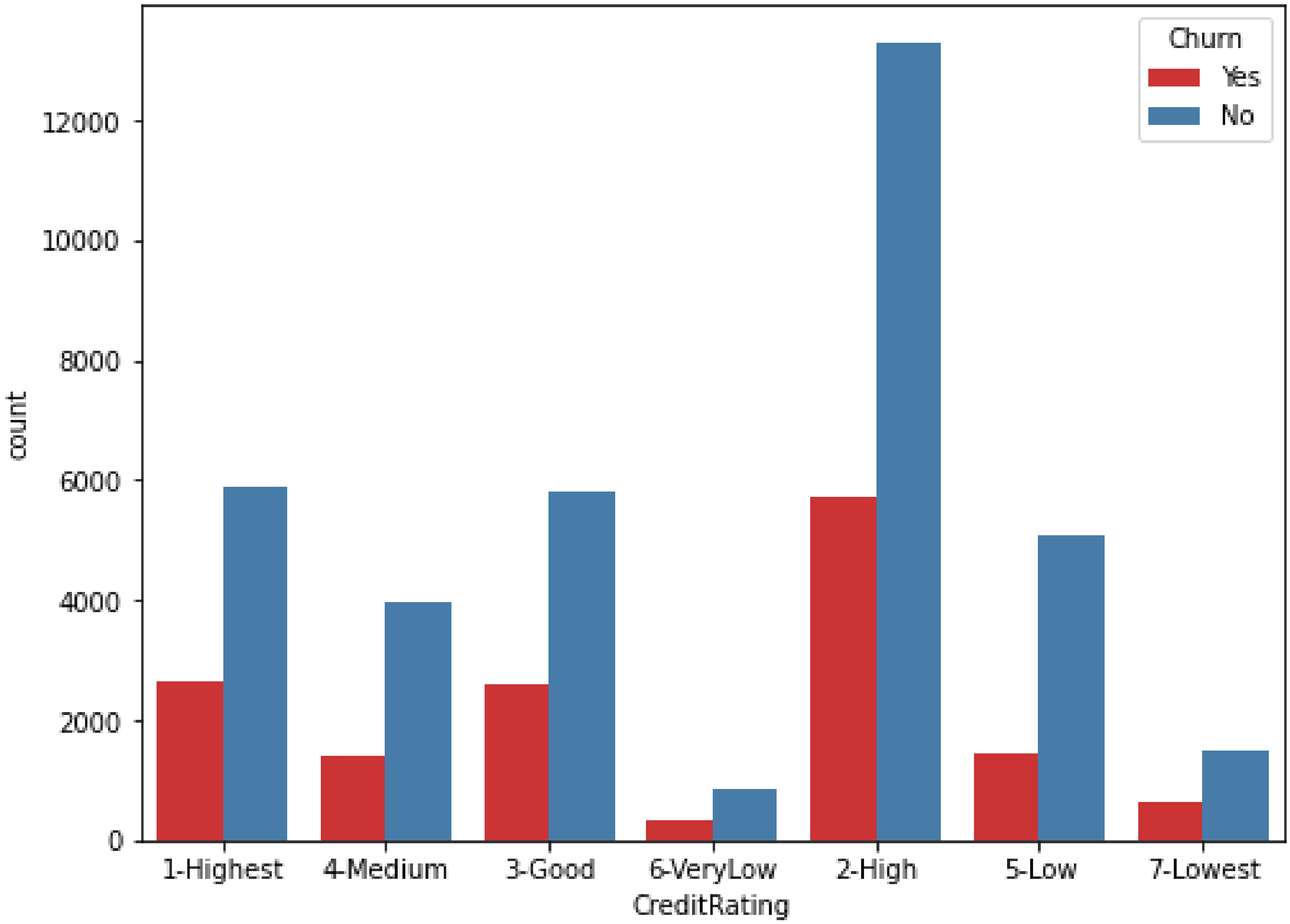
- **Analyze Marital Status column.**



Conclusion: Most of the customers status falls under the unknown category.

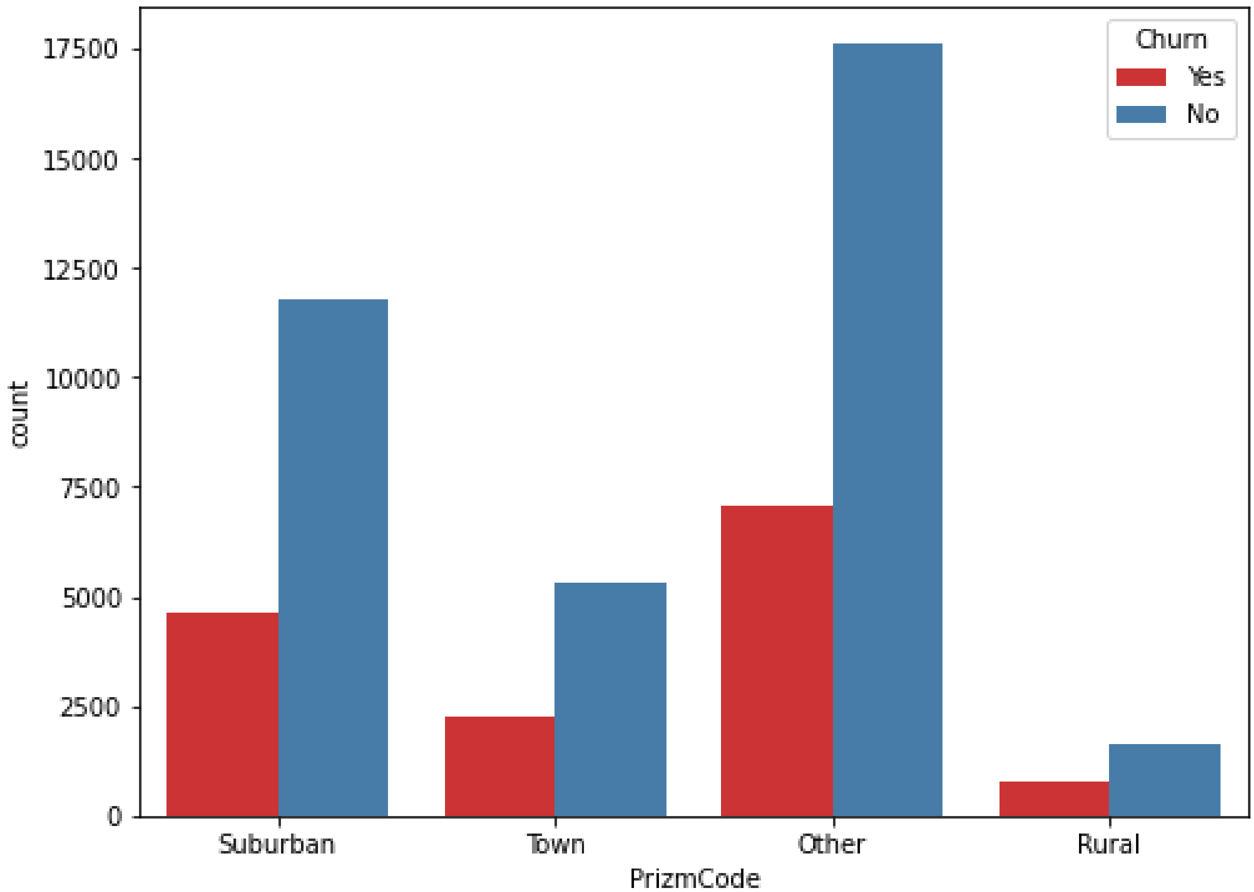
- **Bivariate analysis:**

➢ **Churn vs Credit Rating.**



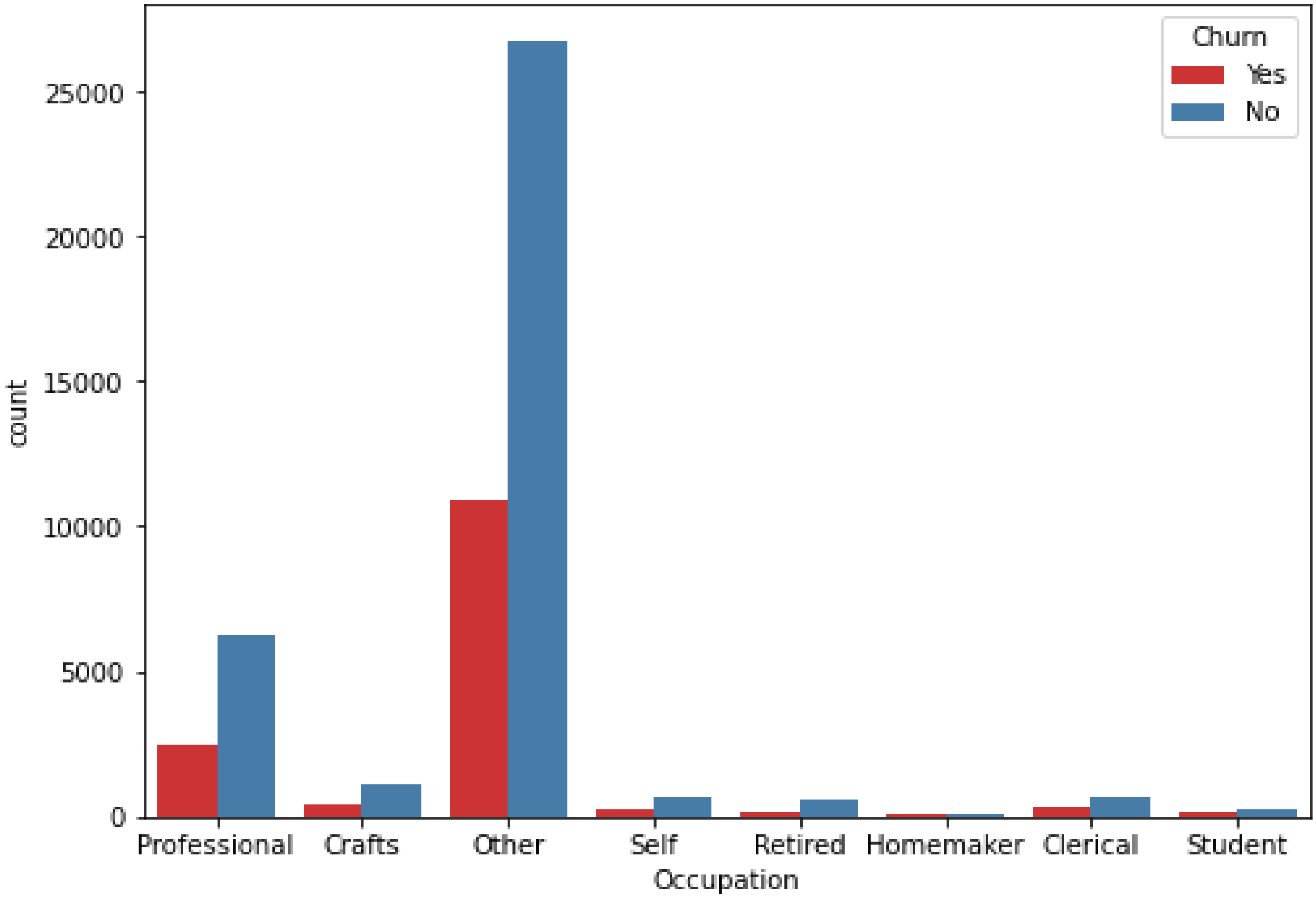
Conclusion: The credit rating of high category has churn rate is high and the very low category churn rate is low.

➢ Churn vs Prizm Code



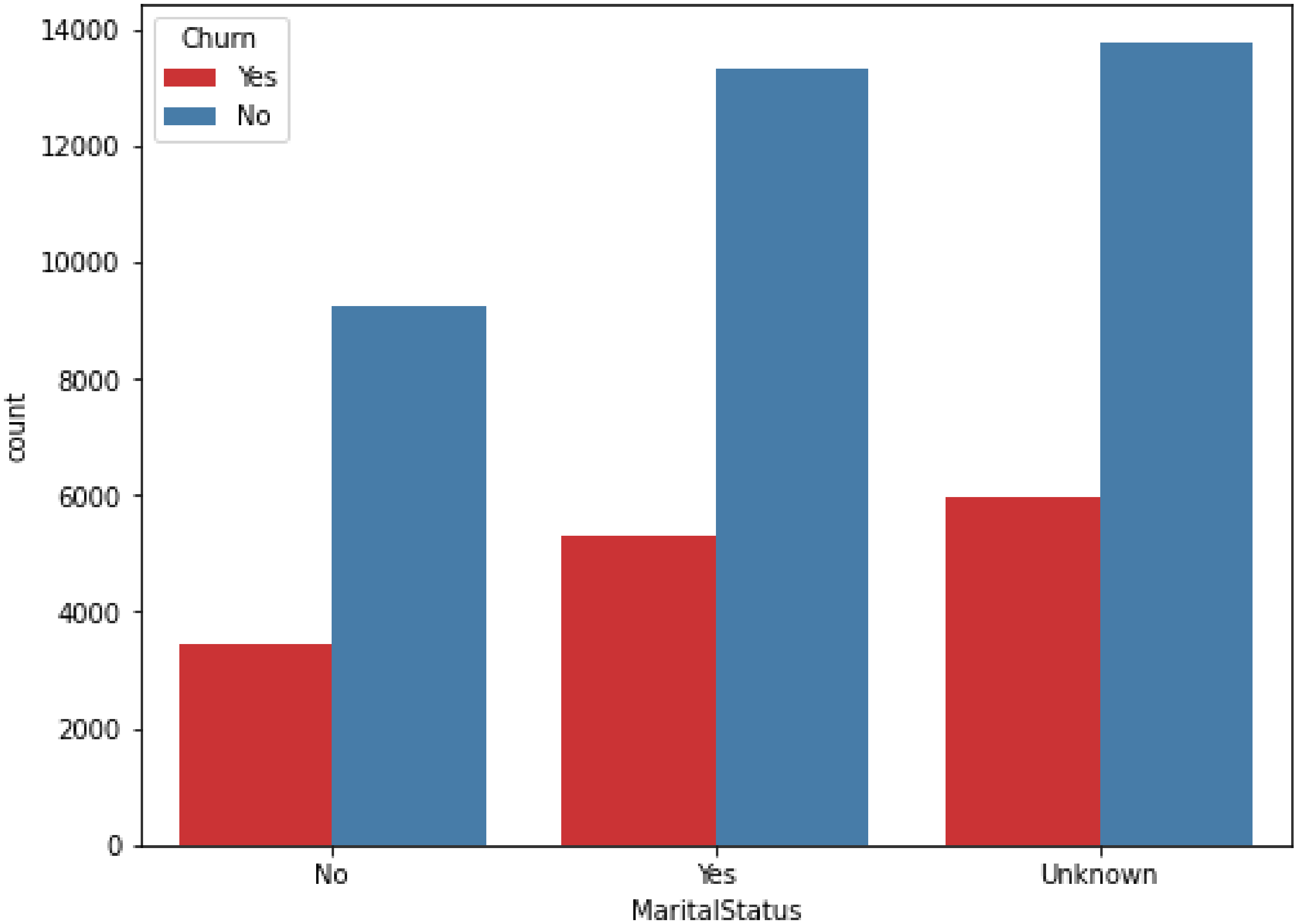
Conclusion: The Prizm code of Other category has churn rate is high and the Rural category churn rate is low.

➢ Churn vs Occupation



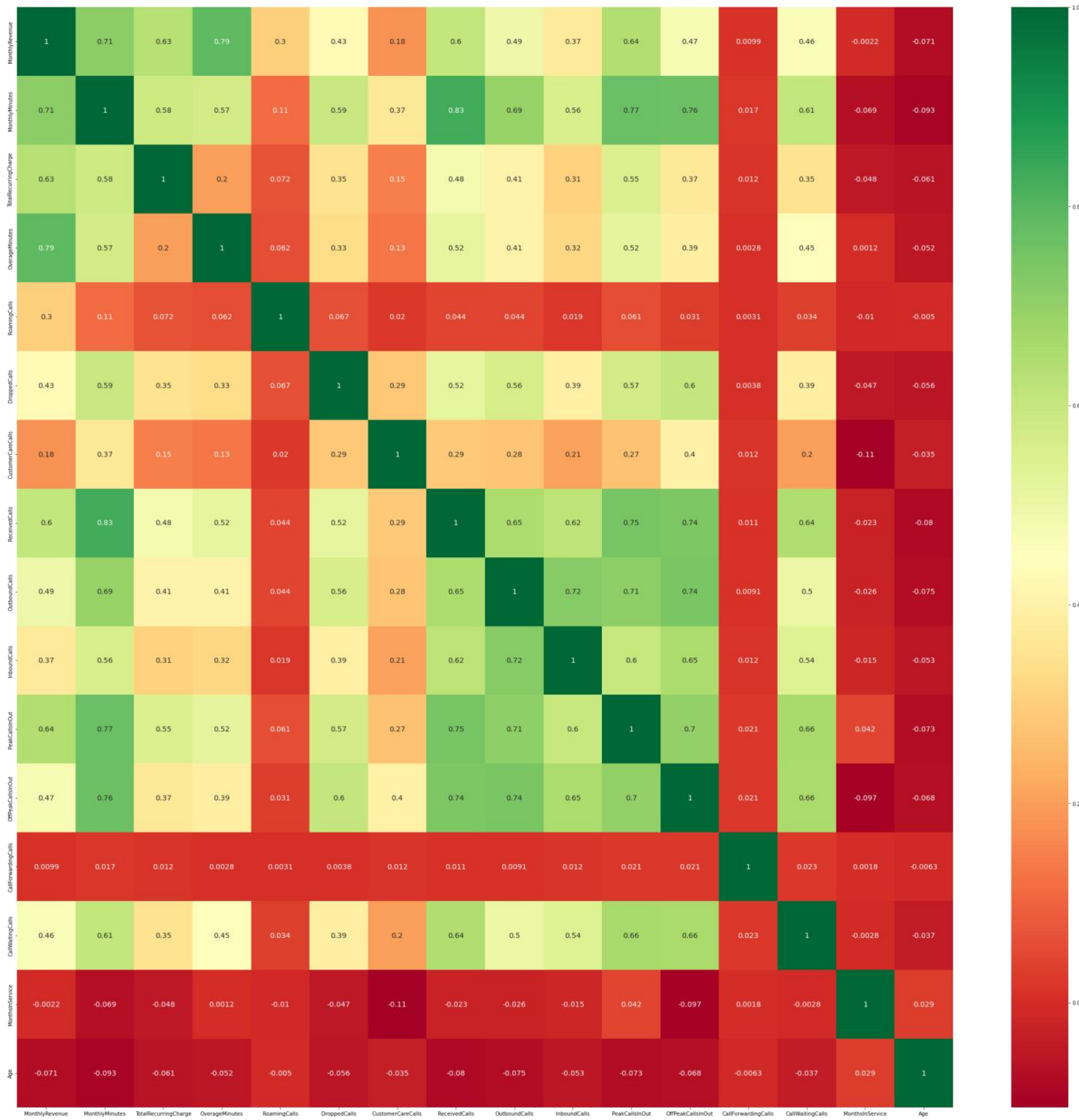
Conclusion: The Occupation of Other category has churn rate is high and the Homemaker category churn rate is low.

- **Churn vs Marital Status**



Conclusion: The Marital Status of Unknown category has churn rate is high and the No category churn rate is low.

➢ Multivariate Data Analysis:



Conclusion: There are many factors that make an employee resign. Using the IBM dataset, some interesting insights were obtained. These insights can be used to build the model.

7. Model Building

1. Train / Test split :

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model. A better option is to split our data into two parts: first one for training our machine learning model, and second one for testing our model.

• Split the dataset into two pieces: a training set and a testing set.

• Train the model on the training set.

• Test the model on the testing set, and evaluate how well our model did.

Advantages of train/test split:
• Model can be trained and tested on different data than the one used for training.

• Response values are known for the test dataset, hence predictions can be evaluated

• Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

8. Model Testing

Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning modelsfacilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.
We have used the following algorithms to build predictive model.

Decision tree: The information gained in the decision tree can be defined as the amount of information improved in the nodes before splitting them for making further decisions.

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.72	0.83	9938
1	0.04	0.45	0.08	272
accuracy			0.71	10210
macro avg	0.51	0.58	0.45	10210
weighted avg	0.95	0.71	0.81	10210

Random Forest: Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

```
Accuracy Score:  0.7205680705190989
Confusion Matrix:
 [[7280 2806]
 [  47   77]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.72      0.84     10086
           1       0.03      0.62      0.05       124

    accuracy                           0.72     10210
   macro avg       0.51      0.67      0.44     10210
weighted avg       0.98      0.72      0.83     10210
```

XGboost Model: Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. **Gradient boosting** refers to a class of ensemble machine learning algorithms. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "*gradient boosting*," as the loss
gradient is minimized.

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.73	0.83	9602
1	0.10	0.46	0.16	608
accuracy			0.71	10210
macro avg	0.53	0.59	0.49	10210
weighted avg	0.90	0.71	0.79	10210

Logistic Regression: Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

```
Accuracy Score:  0.7172380019588639
Confusion Matrix: [[7318 2878]
 [   9    5]]
Classification Report:               precision    recall  f1-score   support

           0       1.00      0.72      0.84     10196
           1       0.00      0.36      0.00        14

    accuracy                           0.72     10210
   macro avg       0.50      0.54      0.42     10210
weighted avg       1.00      0.72      0.83     10210
```

Gradient Boosting Classifier: Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model**.** Decision trees are usually used when doing gradient boosting.

```
              precision    recall  f1-score   support

           0       0.72      0.99      0.84      7327
           1       0.62      0.03      0.06      2883

    accuracy                           0.72     10210
   macro avg       0.67      0.51      0.45     10210
weighted avg       0.69      0.72      0.62     10210
```

Overall Conclusion:

The accuracy of XGBOOST is highest among all the algorithms used to predict customer churn so we can used XGBOOST algorithm for future prediction.

9. Comparing Models

ROC AND AUC:

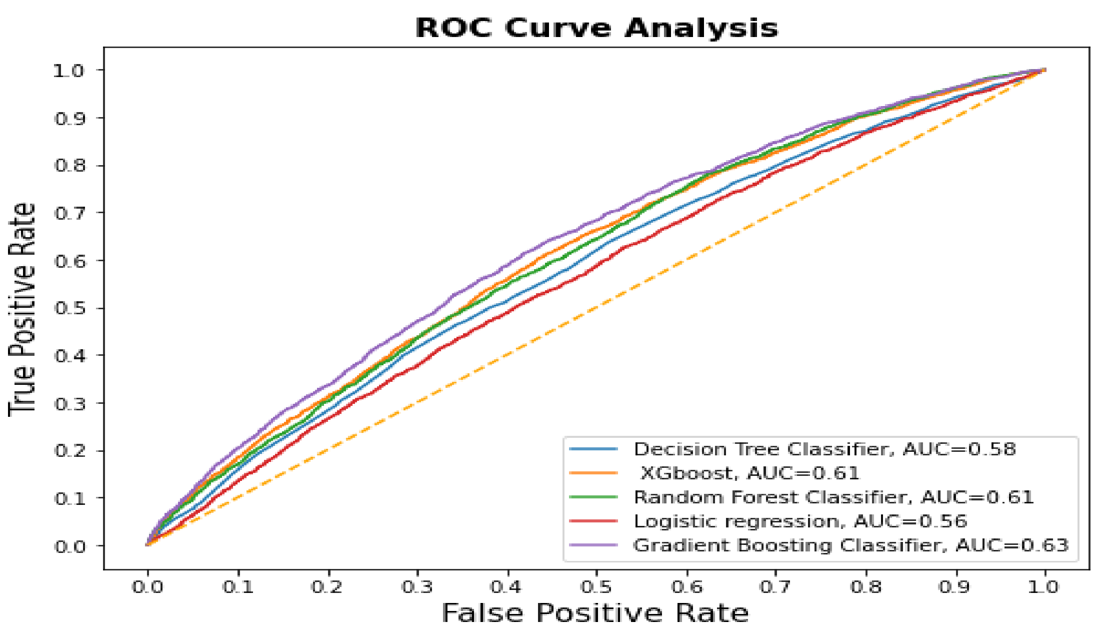
An ROC curve (or receiver operating characteristic curve) is a plot that summarizes the performance of a binary classification model on the positive class.

The x-axis indicates the False Positive Rate and the y-axis indicates the True Positive Rate.

ROC Curve: Plot of False Positive Rate (x) vs. True Positive Rate (y).
The true positive rate is a fraction calculated as the total number of true positive predictions divided by the sum of the true positives and the false negatives (e.g. all examples in the positive class). The true positive rate is referred to as the sensitivity or the recall.

TruePositiveRate = TruePositives / (TruePositives + False Negatives)
The false positive rate is calculated as the total number of false positive predictions divided by the sum of the false positives and true negatives (e.g. all examples in the negative class).
FalsePositiveRate = FalsePositives / (FalsePositives + TrueNegatives)

Multiple ROC-Curves in a single plot:



Conclusion: From above figures we can see that XGboost model has the highest AUC Which is 0.87, So we will use XGboost model for employee attrition datasets. Other models also perform well as there AUC are:

Logistic regression = 0.56
Decision tree classifier = 0.58
XGboost, = 0.61
Gradient boosting classifier = 0.63
Random Forest Classifier = 0.61

10. Conclusion

In our project We developed an efficient and effective approach to analyze Human Resource data, specifically to disclose hidden relationships in our data by drawing behaviors of customer churn from numerous amounts of features available from the data. We built machine learning models to accurately separate churn group from no-churn group and it can be used to predict of any customer who will leave or stay in the service given similar data.

11. Future Scope

Telecom churn prediction is a rapidly evolving field, and there are several potential future directions for research and development. Here are a few examples :

1. Improved data analytics: With the proliferation of big data, there is a need to develop more sophisticated analytical tools and algorithms to handle the vast amounts of data generated by telecom companies. Machine learning techniques such as deep learning and reinforcement learning can help improve the accuracy and reliability of churn prediction models.

2. Integration of new data sources: To improve the accuracy of churn prediction models, telecom companies can consider integrating new data sources such as social media activity, customer feedback, and purchase history. By analyzing these additional data points, telecom companies can gain a more comprehensive view of their customers' behavior and needs.

3. Real-time churn prediction: Real-time churn prediction can help telecom companies take timely action to prevent customer churn. By analyzing customer behavior in real-time, companies can identify early warning signs of churn and take appropriate measures to retain customers.

4. Personalized churn prevention strategies: To retain customers, telecom companies can develop personalized churn prevention strategies based on customer behavior, preferences, and needs. By tailoring retention strategies to individual customers, telecom companies can increase the effectiveness of their retention efforts.

12. Reference

Models:

1. Decision tree:

   https://scikit-learn.org/stable/modules/tree.html

2. Random Forest Regression:

   https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomizedtrees

3. XGBoost regression:

   https://xgboost.readthedocs.io/en/latest/index.html

4. Gradient Boosting Classifier:

   https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d

5. Logistic Regression:

   https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Thank You