Project Report

On

CHURN PREDICTION USING MACHINE LEARNING MODEL

Submitted in partial fulfillment of the requirements for the award of degree of

Bachelor of Technology

in

Computer Engineering (Data Science) - 5th Sem

by

Devansh Jain

(20001016015)

under the supervision of

Ms. Monika Gupta

Department of Computer Engineering

J. C. BOSE UNIVERSITY OF SCIENCE & TECHNOLOGY, YMCA

FARIDABAD-121006

December 2022

CANDIDATE'S DECLARATION

I hereby certify that the work which is being carried out in this Project titled **"Churn Prediction using Machine Learning model"** in fulfillment of the requirement for the degree of Bachelor of Technology in **Computer Engineering (Data Science)** and submitted to "**J. C. Bose University of Science and Technology, YMCA, Faridabad**", is an authentic record of my own work carried out under the supervision of **Ms. Monika Gupta**.

Devansh Jain

20001016015

CERTIFICATE

This is to certify that the work carried out in this project titled **"Churn Prediction using Machine Learning model"** submitted by **Devansh Jain** to "**J. C. Bose University of Science and Technology, YMCA, Faridabad**" for the award of the degree of Bachelor of Technology in **Computer Engineering (Data Science)** is a record of bonafide work carried out by them under my supervision. In my opinion, the submitted report has reached the standards of fulfilling the requirements of the regulations to the degree.

Ms. Monika Gupta
(Mentor)
Assistant Professor
Department of Computer Engg.
J. C. Bose University of Science and Technology, YMCA, Faridabad

Dr. Komal Kumar Bhatia
Chairman,
Department of Computer Engg.,
J. C. Bose University of Science and Technology, YMCA, Faridabad

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

INTRODUCTION

Customer churn is the name for when a subscriber or a regular customer cancels his subscription or stops doing business with a company. Therefore, the churn rate is the measure of how many people stopped being a client of the company in a determined time period.

In business administration, churn is a very important metric of how well the business is doing. If the churn rate is high, then the business is losing a lot of clients and not performing well.

With the evolution of machine learning algorithms and data science, churn prediction has become a very important part of every company's strategy. If a company can accurately predict that a customer is about to churn, it can then act to prevent the churn. Usually working to keep a client is cheaper than working to get a new client.

PROBLEM IDENTIFICATION

Predicting churn is a good way to create proactive marketing campaigns targeted at the customers that are about to churn. Thanks to big data, forecasting customer churn with the help of machine learning is possible. Machine learning and data analysis are powerful ways to identify and predict churn. During churn prediction, you're also:

- Identifying at-risk customers,
- Identifying customer pain points,
- Identifying strategy/methods to lower churn and increase customer retention.

OBJECTIVE

In this project, we'll work with a churn prediction dataset of a phone/internet company from the IBM Developer platform. The main goal is to build a machine learning model capable of accurately predict that a customer will churn based on the information available in the dataset. In order to accomplish that, we'll go through some main steps, such as:

- Exploratory data analysis;
- Data preparation;
- Train, tune, test, and evaluate machine learning models.

CHAPTER 2: DESCRIPTION AND METHODOLOGY

FLOW OF WORK :-

Step 1. Exploratory Data Analysis
Step 2. Data Cleaning (dealing with empty spaces)
Step 3. Analyzing the data (graph plotting)
Step 4. Preparing the data to train the Machine Learning model.
Step 5. Models and Metrics

Exploratory Data Analysis

We'll first import everything we'll need and configure the notebook. Then we'll read the data into a DataFrame and display its first five rows

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	Streami
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	

5 rows × 21 columns

We can see that the first column is an individual key. Also, the majority of the columns contain categorical data.

Data Cleaning

Let's now address the problem in the TotalCharges column. If we ran the code below, and error would be raised informing that some values in this column are filled with whitespaces. Let's then run the next cell and see how many of these values are stored in the column.

Eleven rows present this problem. In the next cell we will:

- Change the whitespaces to null values using np.nan;
- Convert the column to float;
- Use the column's median to fill the null values;
- Check if the values are filled properly and if the values in the column are in the right format.
- Drop the customerID column.

The PaymentMethod column contains four unique values:

- Electronic check
- Mailed check
- Bank transfer (automatic)
- Credit card (automatic)

We'll remove the string (automatic) from the values only so it's easier to understand the charts we'll plot.

Analysing the Data

We'll now plot a bar chart for each categorical column so we can see how each category in these columns impacts the Churn column.

Before plotting, we'll transform the values in the Churn column from labels Yes and No to numbers 0 and 1. We'll do the opposite with the SeniorCitizen column so the column's chart is properly labeled.

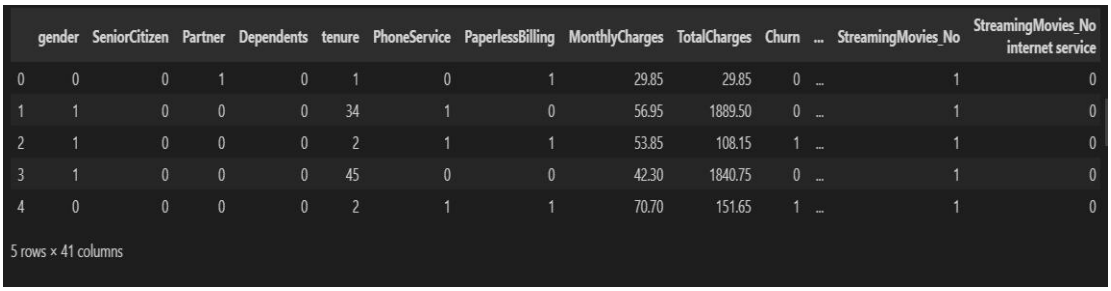
We'll then create three lists:

- One list containing the binary variables, except the Churn;
- One list containing the other categorical variables;
- A third list that will be the addition of the first two.

The third list will be used to plot the charts and the first two we'll be used later on this project to encode the categorical variables.

Preparing the Data

We'll now prepare the data to train the models. First, we need to encode the categorical variables. For the binary variables, we'll use the LabelEncoder class and for the other categorical variables we'll use the pandas.get_dummies function.



	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges	Churn	...	StreamingMovies_No	StreamingMovies_No internet service
0	0	0	1	0	1	0	1	29.85	29.85	0	...	1	0
1	1	0	0	0	34	1	0	56.95	1889.50	0	...	1	0
2	1	0	0	0	2	1	1	53.85	108.15	1	...	1	0
3	1	0	0	0	45	0	0	42.30	1840.75	0	...	1	0
4	0	0	0	0	2	1	1	70.70	151.65	1	...	1	0

5 rows × 41 columns

Before training the models, we need to split the data between train and test. In the next cell, we'll also use the RandomUnderSampler class to balance the dataset. We'll then have both balanced and unbalanced training sets.

Models and Metrics

We'll use four different machines learning algorithms:

- Logistic Regression;
- Decision Trees;
- Support Vector Machines - SVM;
- XGBoost.

Each of the algorithms we'll be trained using both balanced and unbalanced data so we can see which algorithm-data combination yields the bests results.

Metrics

The most important metric we'll use is Recall. This metric indicates the proportion of positive results yielded by the model by the total number of positive labels in the dataset. In this case, the Recall reveals the proportion of churns identified correctly by the total number of churns.

Recall is calculated using the following equation:

$$TP/(TP+FN)$$

Where:

- TP: True Positives;
- FN: False Negatives.

We'll also keep track of Precision as a secondary metric. Precision indicates the proportion of positives yielded by the models that are actually true positives.

For the problem we are dealing with, Recall is more important because it's preferable to have a model that does not miss any churns but sometimes classify a non-churns as churns, than a model that does not classify non-churns as churns but misses a lot of churns. In other words, we rather be incorrect when classifying a non-churning costumer than when classifying a churning customer.

We'll now train all the models using cross-validation and store all the results in a DataFrame so we can better visualize them.

Hyperparameters Tunning

We can see from the DataFrame above that the models that used balanced data yielded better results. So we'll stick with that balanced data.

SVM, Logistic Regression, and XGBoost provided similar results in both metrics. Therefore, we'll tune hyperparameters using grid search in all these three models. For that, we'll only considerer Recall for evaluation.

CHAPTER 3: HARDWARE AND SOFTWARE REQUIREMENTS

Hardware and software requirements of any project are must to be satisfied, so that the virtual environment can be set up on any machine to run the project. So, in this section the software and the hardware requirements are discussed completely.

Software Requirements

Code Editor – VS Code

OS – Windows 10

Language – Python

Libraries – Pandas, NumPy, Seaborn, Matplotlib, Skicit-learn

Hardware Requirements

a. A computer
b. Minimum System Requirements: -
    i. O.S. -Windows 10
    ii. Quad-core i5 CPU minimum
    iii. 8 GB RAM
    iv. 256GB SSD

CHAPTER 4: RESULTS AND ANALYSIS

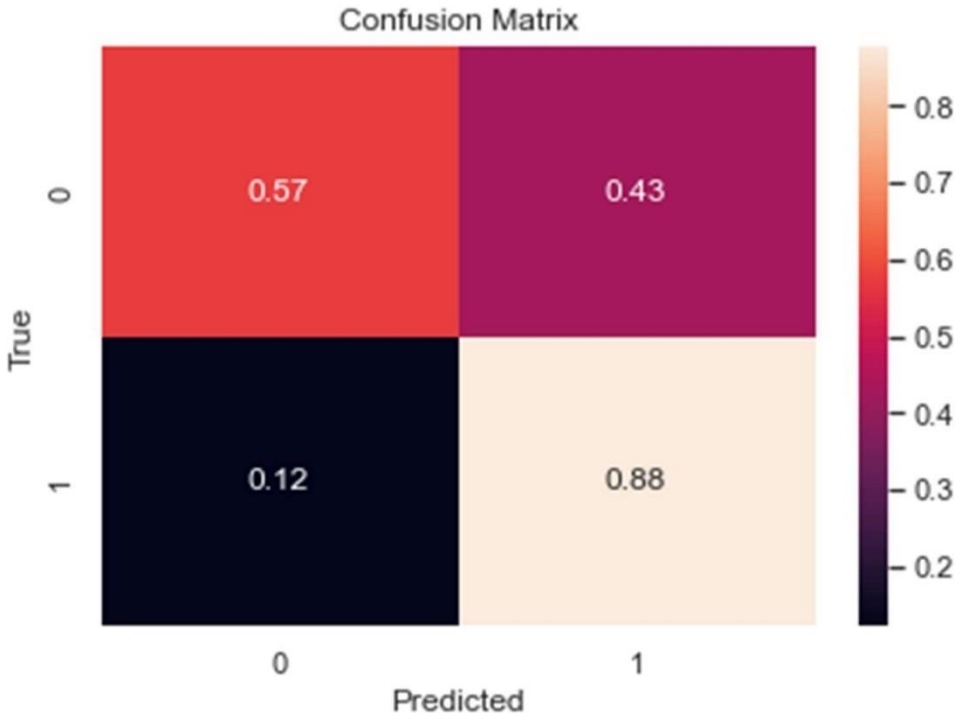
We can learn a lot from these charts. Here are some insights:

- Customers without dependents are two times more likely to churn.
- Customers that use paperless billing and optical fiber are more likely to churn.
- Customers with no online security or backup, no device protection, and no tech support are from two to three times more likely to churn.
- Customers with no internet service are unlikely to churn.
- Customers with month-to-month contracts are almost four times more likely to churn than customers with yearly contracts. Two-year contractors are very unlikely to churn.
- Customers that use electronic checks to pay their bills are more likely to churn.

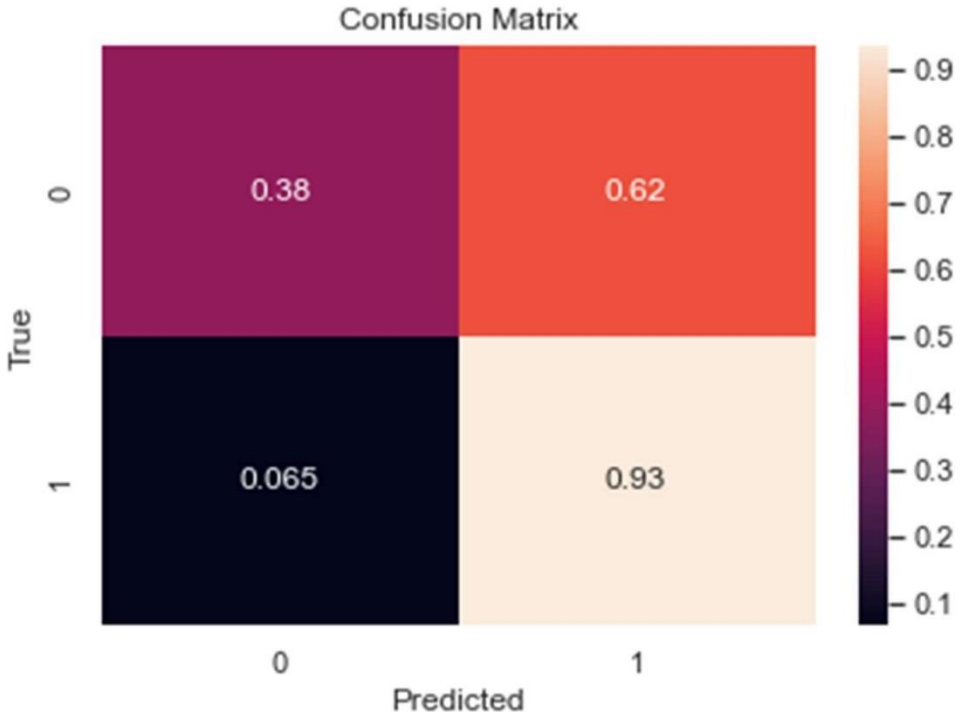
We can see there is a significant correlation between
the tenure and Churn columns. The highest the tenure, the lowest the chances that the customer will churn.

Unfortunately, there's no much we can say about
the MonthlyCharges and TotalCharges columns.

By XGBoost model :-



By SVM :-



Both models performed as well on the test set as on the training set. But SVM is still presenting a better Recall, therefore it would be the model chosen in a real-life situation.

CHAPTER 5: CONCLUSION

EXPECTED OUTCOMES

In this project, we worked on a churn prediction problem where the main goal was to build a machine learning model capable of correctly identify the highest possible number of churning customers. For that, we went through the following steps:

- Exploratory data analysis;
- Data cleaning
- Data analysis
- Data preparation
- Training, tunning, and evaluating machine learning models.

As a result, we have two models that presented very satisfactory outcomes. We can then consider that objective of the project was accomplished.

BRIEF PROFILE OF STUDENT

Name : Devansh Jain

Roll No : 20001016015

Branch : Computer Engineering (Data Science)

Email Id : 20001016077@jcboseust.ac.in

Brief About Project:

The Project revolves around building a machine learning model capable of accurately predict that a customer will churn based on the information available in the dataset.

Machine learning and data analysis are powerful ways to identify and predict churn. During churn prediction, you're also:

- Identifying at-risk customers,
- Identifying customer pain points,
- Identifying strategy/methods to lower churn and increase customer retention.

Future Scope:

The project includes prediction of churn rate only and tells us which machine learning model provides us the best result. We can further build a model that can take decisions itself, inform the users about increasing churn rate and also tell the users possible solutions to lower the churn for the betterment of their organization.