

TELECOM CUSTOMER CHURN PREDICTION USING IBM WATSON

1. INTRODUCTION

1.1 Overview

Customer churn has become highly important for companies because of increasing competition among companies, increased importance of marketing strategies and conscious behaviour of customers in recent years. Customers can easily trend toward alternative services. Companies must develop various strategies to prevent these possible trends, depending on the services they provide. During the estimation of possible churns, data from the previous churns might be used. An efficient churn predictive model benefits companies in many ways. Early identification of customers likely to leave may help to build cost effective ways in marketing strategies.

Telecommunication industry always suffers from a very high churn rates when one industry offers a better plan than the previous there is a high possibility of the customer churning from the present due to a better plan in such a scenario it is very difficult to avoid losses but through prediction we can keep it to a minimal level. A machine learning model is built and this helps to identify the probable churn customers and then makes the necessary business decisions.

1.2 Purpose

Customer churn is a common problem across businesses in many sectors. If you want to grow as a company, you have to invest in acquiring new clients. Every time a client leaves, it represents a significant investment lost. Both time and effort need to be channel led into replacing them. Being able to predict when a client is likely to leave, and offer them incentives to stay, can offer huge savings to a business.

As a result, understanding what keeps customers engaged is extremely valuable knowledge, as it can help you to develop your retention strategies, and to roll out operational practices aimed at keeping customers from walking out the door.

Predicting churn is a fact of life for any subscription business, and even slight fluctuations in churn can have a significant impact on your bottom line. We need to know: “Is this customer going to leave us within X months?” Yes or No? It is a binary classification task.

2. LITERATURE SURVEY

2.1 Existing problem

To predict the telecom customers who are likely to exit the contract and generate patterns of Churn and nonchurn to assist the management in making appropriate decisions to limit churn. Most telecom companies suffer from voluntary churn. The churn rate has a strong impact on the customer's lifetime value because it affects the length of service and the future revenue of the company. It is estimated that 75 percent of the 17 to 20 million subscribers signing up with a new wireless carrier every year are coming from another wireless provider, which means they are churners.

Telecom companies are battling to attract each other's customers while retaining their own. Thus, Customer churn reduction is the central concern of most telecom companies as switching costs to the customer are low

and acquisition cost to the company is high. Churn reduces profitability as it means potential loss of future revenue and also losing the invested costs of acquisition.

Ways to Reduce Customer Churn:

Consistently Exceed Customers' Expectations

The most fundamental way to decrease your churn rate is by keeping your customers happy. While you definitely want to avoid letting them down, you have to look for areas to go over and above your customer's expectations and delight them. Failing to deliver on a promise is one of the fastest ways to lose a customer, and many companies say that dissatisfaction and unmet expectations are among the top reasons for client churn.

Provide Awesome Customer Service This one should go without saying, but if you've ever spent half an hour listening to the hold muzak waiting for a disinterested, incompetent customer service rep to "assist you," you'll know that some companies simply don't put enough effort into customer service.

Create Switching Costs:

Switching costs are any cost that a customer incurs by trading one product or service for another. Higher switching costs naturally reduce churn by reducing the likelihood that a customer will switch to a substitute product instead of returning to your brand.

2.2 Proposed solution

The main objective of this paper is to predict whether the customer will churn or not. And to prevent customers from switching necessary options that, as extending the packages or offering new incentives or services for those customers who are likely to switch to another service provider from the current service provider [5]. Because these services which are provided by the company mainly generate revenue. Data mining techniques have been broadly used to develop a model of churn prediction. In this paper, we use a classification algorithm with an optimization algorithm to find out the optimized customers

Dataset

Telecommunication data faces a large number of strict policies and rules that make access to data extremely difficult. For these reasons, we used the dataset which was obtained from the www.kaggle.com and it contains 21 attributes and 3333 instances Methods.

Algorithm

Input: Telecom churns dataset Output: Optimized churners from the total churners Step 1: Apply the classification Naïve Bayesian Algorithm. Step 2: Compare the posterior probability of classes and the highest one is chosen as the predicted output. Step 3: Apply Elephant Herding Optimization.

Logistic Regression:

Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). o We use ROC (Receiver Operating Characteristics) curve to set a threshold value that minimizes the false positive rate and maximizes the true positive rate.

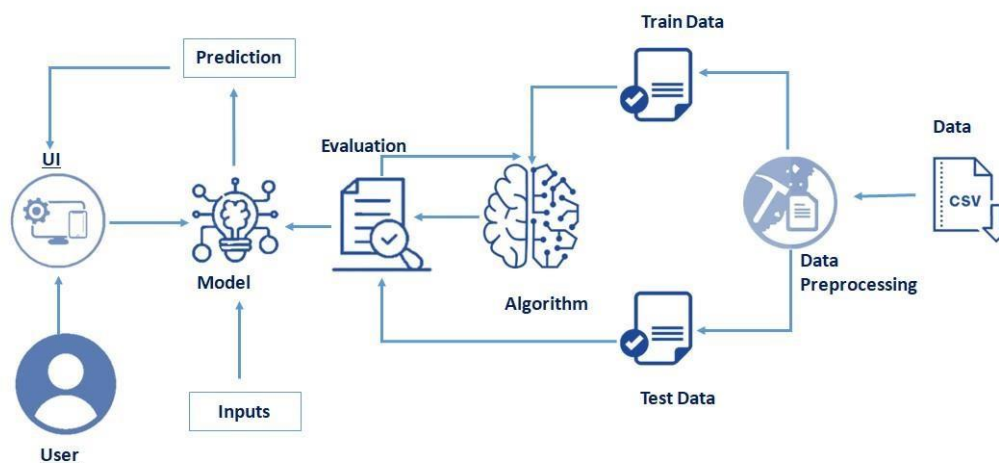
Decision Trees:

A decision tree is a type of supervised learning algorithm (having a pre-defined target Variable) that is mostly used in classification problems. It works for both categorical and a continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets using the greedy approach.

- o Gini Index – (Classification and Regression Trees).

3. THEORETICAL ANALYSIS

3.1 Block diagram



3.2 Software Requirement/Hardware Requirements

Packages -

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Flask
6. Imblearn

Hardware requirement

Processor: intel Core i3

Hard Disk space: min 400GB

Ram: 4GB

4.EXPERIMENTAL INVESTIGATIONS

How to predict and prevent customer churn has become a focus that many companies and scholars are concerning. As a result of the automation of operation flow, the enterprises have accumulated plenty of business data during their daily operating activities, which gives the data mining technology a good basis to work at. In the past decades, lots of algorithms and models have been used in this field and some scholars have worked on the comparison among different methods. Actually, there is no method that can be better than others in all indicators, because accuracy and concision can't appear in one method simultaneously. Until today, many algorithms and models have been used in predicting customer churn. Some models such as Decision tree, Artificial Neural Network, and Logistic regression have been used frequently and some other models such as Bayesian Network, Support Vector Machine [6], Rough Set, and Survival Analysis are less.

3.3 Data Acquisition

Data Acquisition is a very important and difficult task for a churn prediction model in telecom. Any telecom company will not provide a database to the public, this is because the private information of customers cannot be misused. However, some telecom companies' database is publicly available on data repository websites. The database used in this study is from an American telecom company named Orange. This customer's periodic data contains the customer's behavior data in that particular period. This database contains 3333 customer information and has 20 attributes. "Churn" is the independent variable representing the status of the customer churn or non-churn. This database has 483 churners. This usage behavior of customers will be used to predict the behavior of the near future.

3.4 Data preparation

In any churn prediction, model data preparation is essential and time taking task. In this phase, data are gathered, integrated, and cleaned. In the integrated phase, data is collected from different resources and get into the required form. In data, cleaning any ambiguity, errors, and unnecessary data is removed. In the telecom database, not all fields are used for the prediction. Some fields that contain text data are removed like state, DOB, etc. Fields with so many null values are also removed from the database. In addition, remove two fields that have the same meaning for prediction are removed, as that features will only cause in increasing the database size. A database with too many attributes decreases the model speed. This database had 20 attributes at the initial and after this phase 11 attributes were retained in the database. Then the database was converted to a binary form for the experiment.

5. RESULT



Telecom Customer Churn Predict x Telecom Customer Churn Predict x +

127.0.0.1:5000/assessment?

Prediction form

Gender **Gender**
Senior Citizen **Senior Citizen**
Partner **Partner**
Dependents **Dependents**
Tenure
Phone Services **Phone Services**
Multiple Lines **Multiple Lines**
Internet services **Internet services**
Online Services **Online Services**
Online Backup **Online Backup**
Device Protection **Device Protection**
Tech Support **Tech Support**
Streaming TV **Streaming TV**
Streaming Movies **Streaming Movies**
Contract **Contract**
Paperless Billing **Paperless Billing**
Payment Methods **Payment Methods**
Monthly Charges
Total Charges
Submit

Activate Windows
Go to Settings to activate Windows.

Type here to search

ENG IN 7:35 PM 10/13/2022

Telecom Customer Churn Predict x Telecom Customer Churn Predict x +

127.0.0.1:5000/assessment?

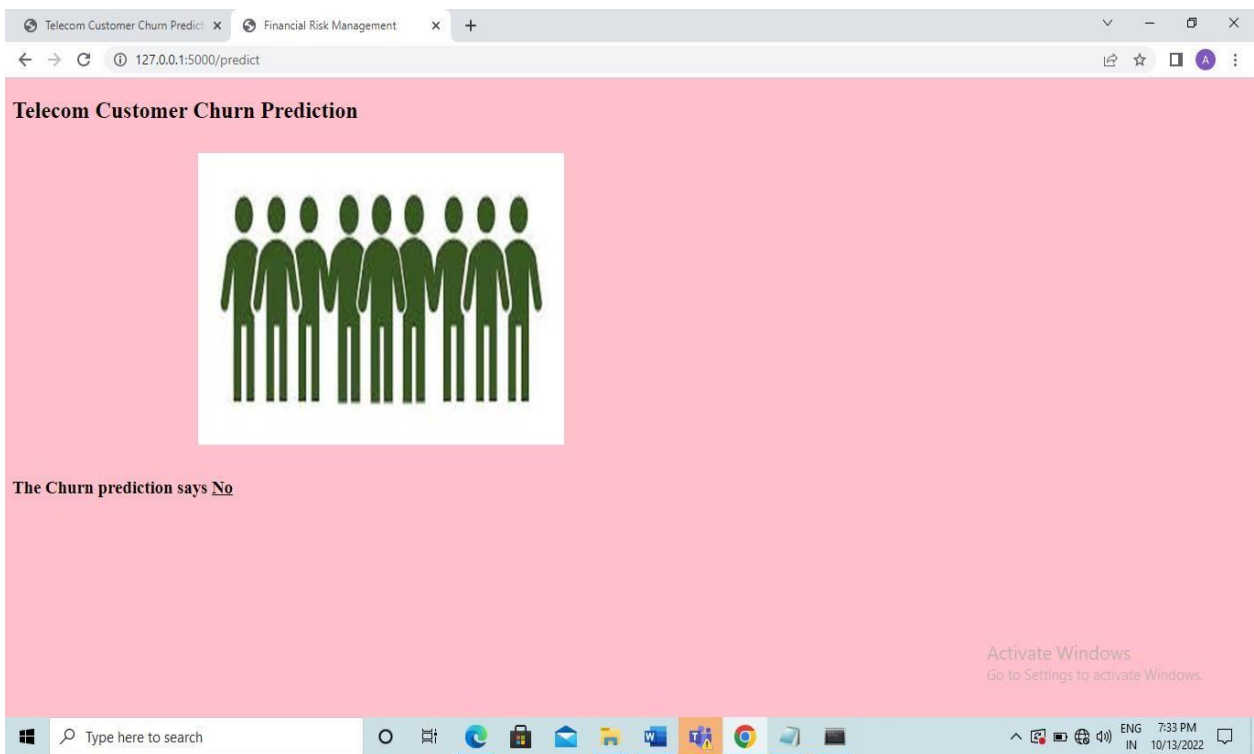
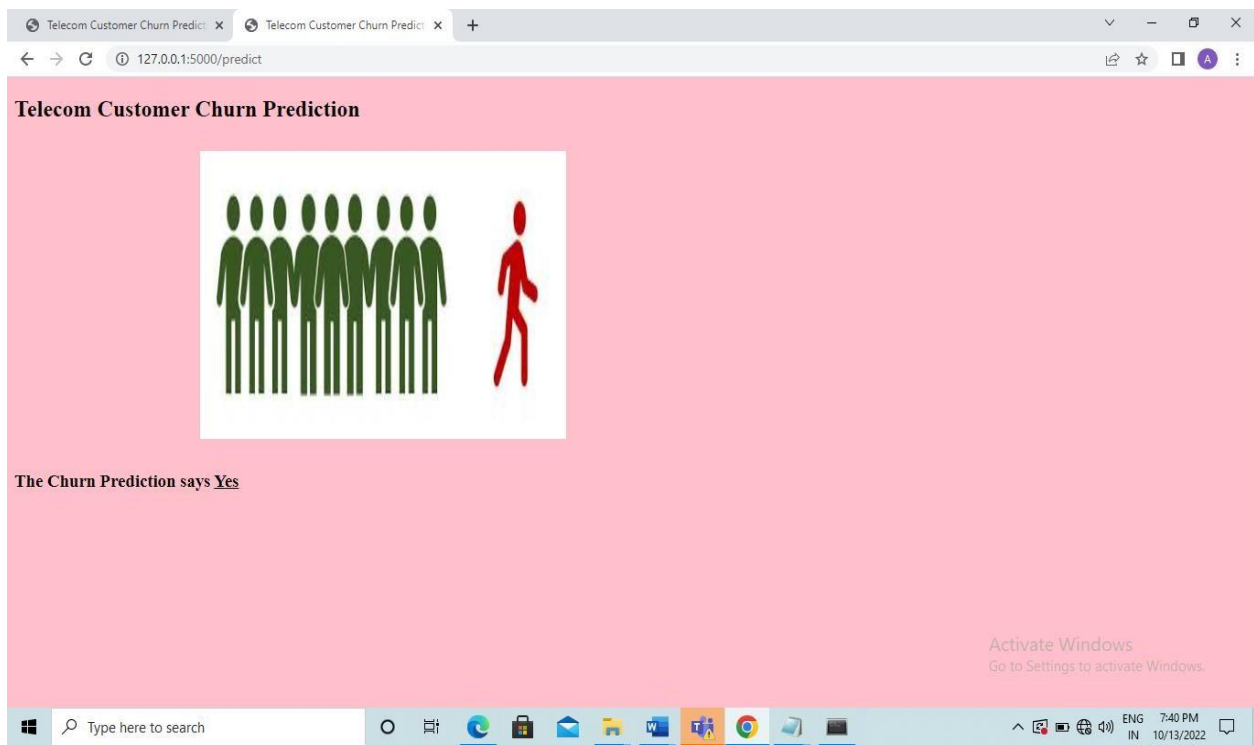
Prediction form

Male **Male**
Yes **Yes**
Yes **Yes**
Yes **Yes**
34
Yes **Yes**
Yes **Yes**
Fibre Optics **Fibre Optics**
Yes **Yes**
Yes **Yes**
No **No**
Yes **Yes**
Yes **Yes**
Yes **Yes**
One year **One year**
Yes **Yes**
Mail Check **Mail Check**
54.32
52.69
Submit

Activate Windows
Go to Settings to activate Windows.

Type here to search

ENG IN 7:34 PM 10/13/2022



6. ADVANATGES AND DISADVANTGES

6.1 Advantages

1. Identify at-risk customers

For any business that wants to enjoy the benefits of customer churn prediction, machine learning opens dozens of opportunities. Machine learning is able to analyze client behavior and measure their probability of churning. In particular, to precisely identify churn rate, machine learning algorithms can be trained to learn the behavior patterns of clients/partners who have already canceled their contracts or any other relationships with a particular company and compare them with the existing ones. Then correlations between the actions of active and inactive clients are done. As a result, the algorithm recognizes the customers that are more likely to leave.

2. Identify pain points

Different companies lose their clients for different reasons. In most cases, there are numerous "pain points," which remain unknown for product owners. From the bad quality and absent features to unpleasant design and poor customer service — there are a lot of details which you do not take into account that your clients do. Even if your product is almost perfect, you can still reward your new customers with some attractive discounts and offers and ignore your loyal ones. When a business applies churn prediction, machine learning can do analysis and forecasts based not only on customer behavior but also on the brand's.

3. Identify methods to implement

After the root cause of client churn has been identified, companies can reconsider and rebuild their products and change their business strategy accordingly. Transformed data and automated flow can be used in CRM and marketing automation systems. However, this doesn't mean that using machine learning for churn prediction is about building a certain model for a certain task. It is more about domain knowledge and an ability to deliver the best possible solution based on learning data, processes, and behavior.

6.2 Disadvantages

1. Doesn't provide clarity on the types of customers leaving which means you couldn't find out which one left, the new or old customer.
2. Doesn't differentiate companies between industry types.

7. APPLICATIONS

In this competitive world, business becomes highly saturated. Especially, the field of telecommunication faces complex challenges due to a number of vibrant competitive service providers. Therefore it has become very

difficult for them to retain existing customers. Since the cost of acquiring new customers is much higher than the cost of retaining the existing customers, it is the time for the telecom industries to take necessary steps to retain the customers to stabilize their market value. This paper explores the application of data mining techniques in predicting likely churners and the impact of attribute selection on identifying the churn. It also compares the efficiency of Decision tree and Neural Network classifiers and lists their performances.

8. CONCLUSION

The importance of this type of research in the telecom market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income to telecom companies. Hence, this research aimed to build a system that predicts the churn of customers in SyriaTel telecom company. These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 70% for training and 30% for testing. We chose to perform cross-validation with 10-folds for validation and hyperparameter optimization. We have applied feature engineering, effective feature transformation and selection approach to make the features ready for machine learning algorithms. In addition, we encountered another problem: the data was not balanced. Only about 5% of the entries represent customers' churn. This problem was solved by undersampling or using tree algorithms not affected by this problem. Four tree based algorithms were chosen because of their diversity and applicability in this type of prediction. These algorithms are Decision Tree, Random Forest, GBM tree algorithm, and XGBOOST algorithm.

9. FUTURE SCOPE

The future scope of this paper will use hybrid classification techniques to point out existing association between churn prediction and customer lifetime value. The retention policies need to be 75.00% 80.00% 85.00% 90.00% 95.00% 100.00% 105.00% Small Data Set Large Data Set Decision Tree Logistic Regression considered by selecting appropriate variables from the dataset. The passive and the dynamic nature of the industry ensure that data mining has become increasingly significant aspect in the telecommunication industry prospect.

10. BIBLIOGRAPHY

Installation of Anaconda Navigator:

<https://www.youtube.com/embed/5mDYijMfSzs>

Installation Of Python Packages:

https://www.youtube.com/embed/akj3_wTploU Data

Collection:

<https://www.kaggle.com/shrutimechlearn/churn-modelling>

Data Pre-processing:

Data Pre-processing includes the following main tasks

1. Import the Libraries.
2. Reading the dataset.
3. Exploratory Data Analysis
4. Checking for Null Values.
5. Data Visualization.
6. Label Encoding
7. OneHot Encoding
8. Splitting the Dataset into Dependent and Independent variables.
9. Splitting Data into Train and Test

Handling Null Values:

<https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e>

Data Visualization:

<https://www.youtube.com/embed/TLdXM0A7SR8>

Splitting Dependent And Independent Columns:

https://www.youtube.com/embed/A_V6daPQZIU Splitting The

Data Into Train And Test:

<https://www.youtube.com/embed/xgDs0scjuuQ> Training And

Testing The Model:

<https://www.youtube.com/embed/yIYKR4sgzI8>

Model Evaluation:

<https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226>

Flask Frame Work Reference:

https://www.youtube.com/embed/lj4I_CvBnt0

Flask Refarance To Run:

<https://www.youtube.com/embed/UbCWoMf80PY>

Train The Model On IBM:

Account Creation:

<https://cloud.ibm.com/registration>

Train Model On IBM Watson:

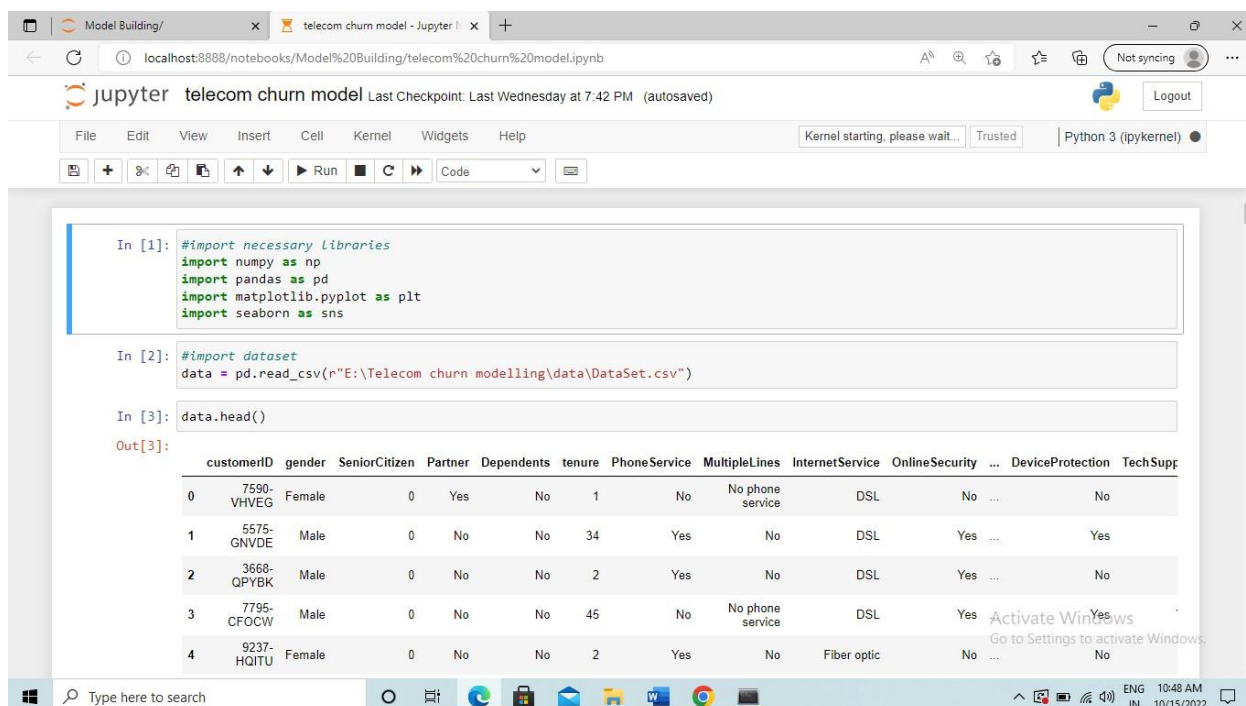
<https://youtu.be/TysuP3KgSzc>

Integrate Flask With Scoring Endpoint:

<https://www.youtube.com/embed/ST1ZYLmYw2U>

APPENDIX

Python source code



The screenshot shows a Jupyter Notebook titled "telecom churn model" running on a local host. The code in the notebook is as follows:

```
In [1]: #import necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

In [2]: #import dataset
data = pd.read_csv(r"E:\Telecom churn modelling\data\DataSet.csv")

In [3]: data.head()
```

The output of the third cell shows the first five rows of the dataset:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupp
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3568-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

Model Building/ x telecom churn model - Jupyter x +

localhost:8888/notebooks/Model%20Building/telecom%20churn%20model.ipynb

jupyter telecom churn model Last Checkpoint: Last Wednesday at 7:42 PM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [4]: `data['Churn'].value_counts() # Data is imbalanced`

Out[4]:

```
No    5174
Yes   1869
Name: Churn, dtype: int64
```

In [5]: `data.drop(["customerID"], axis=1, inplace=True)`

In [6]: `data.head()`

Out[6]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
0	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	
1	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	
2	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	
3	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	
4	Female	0	No	No	2	Yes	No	Fiber optic	No	No		

Windows taskbar: Type here to search, 10:49 AM, 10/15/2022

Model Building/ x telecom churn model - Jupyter x +

localhost:8888/notebooks/Model%20Building/telecom%20churn%20model.ipynb

jupyter telecom churn model Last Checkpoint: Last Wednesday at 7:42 PM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Churn
dtype: int64

In [13]: `data.corr()`

Out[13]:

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
SeniorCitizen	1.000000	0.016567	0.220173	0.102652
tenure	0.016567	1.000000	0.247900	0.825464
MonthlyCharges	0.220173	0.247900	1.000000	0.650864
TotalCharges	0.102652	0.825464	0.650864	1.000000

In [14]: `sns.heatmap(data.corr(), annot=True)`

Out[14]: <AxesSubplot:>

Windows taskbar: Type here to search, 10:49 AM, 10/15/2022

Model Building/ x telecom churn model - Jupyter i x +

localhost:8888/notebooks/Model%20Building/telecom%20churn%20model.ipynb

jupyter telecom churn model Last Checkpoint: Last Wednesday at 7:42 PM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

one.fit_transform(x_resample[:,13:14]).toarray() i= one.fit_transform(x_resample[:,14:15]).toarray() j= one.fit_transform(x_resample[:,16:17]).toarray()
x_resample=np.delete(x_resample[[6,7,8,9,10,11,12,13,14,16],axis=1) x_resample=np.concatenate((a,b,c,d,e,f,g,h,i,j,x_resample),axis=1)

In [31]: `from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x_resample,y_resample,test_size = 0.2, random_state = 0)`

In [32]: `from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.fit_transform(x_test)`

In [33]: `from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(random_state=0)
lr.fit(x_train,y_train)`

Out[33]: `LogisticRegression
LogisticRegression(random_state=0)`

In [34]: `lr_pred = lr.predict(x_test)`

In [35]: `lr_pred`

Out[35]: `array([1, 0, 1, ..., 0, 0, 1])`

Windows taskbar: Type here to search, 10:50 AM, 10/15/2022

Model Building/ x telecom churn model - Jupyter i x +

localhost:8888/notebooks/Model%20Building/telecom%20churn%20model.ipynb

jupyter telecom churn model Last Checkpoint: Last Wednesday at 7:42 PM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Out[45]: `0.7652173913043478`

In [46]: `dtc_cm = confusion_matrix(dtc_pred,y_test)`

In [47]: `dtc_cm`

Out[47]: `array([[621, 74],
[412, 963]], dtype=int64)`

In [48]: `from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators = 10, criterion = "entropy", random_state=0)
rfc.fit(x_train,y_train)`

Out[48]: `RandomForestClassifier
RandomForestClassifier(criterion='entropy', n_estimators=10, random_state=0)`

In [49]: `rfc_pred = rfc.predict(x_test)`

In [50]: `rfc_pred`

Out[50]: `array([1, 1, 1, ..., 1, 1, 1])`

In [51]: `rfc_acc = accuracy_score(rfc_pred,y_test)`

Windows taskbar: Type here to search, 10:50 AM, 10/15/2022

Spyder (Python 3.9)

File Edit Search Source Run Debug Consoles Projects Tools View Help

C:\Users\ACER

E:\Telecom churn modelling\IBM\flask\apbm.py

temp.py x scoring_endpoint.py x apbm.py x

```
1 from flask import Flask, render_template, request
2 app = Flask(__name__)
3 #import pickle
4 #model = pickle.load(open('churn.pkl', 'rb'))
5 import requests
6
7 # NOTE: you must manually set API_KEY below using information retrieved from your
8 API_KEY = "iMzFyOH5yWpWzGz5KMCL8H9sPLV98EhvNuDLG9HHox0"
9 token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={
10     'API_KEY': 'grant_type: urn:ibm:params:oauth:grant-type:apikey'})
11 mltoken = token_response.json()['access_token']
12
13 header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}
14
15
16 @app.route('/')
17 def helloworld():
18     return render_template("base.html")
19 @app.route('/assessment')
20 def prediction():
21     return render_template("index.html")
22
23 @app.route('/predict', methods = ['POST'])
24 def admin():
25     a= request.form["gender"]
26     if (a == 'f'):
27         a=0
28     if (a == 'm'):
29         a=1
30     b= request.form["srcitizen"]
31     if (b == 'n'):
32         b=0
33     if (b == 'y'):
34         b=1
35     c= request.form["partner"]
```

Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Help.

New to Spyder? Read our [tutorial](#)

Help Variable Explorer Plots Files

Console 1/A x

Python 3.9.12 (main, Apr 4 2022, 05:22:27) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 8.2.0 -- An enhanced Interactive Python.

In [1]:

Activate Windows
Go to Settings to activate Windows.

LSP Python: ready conda: base (Python 3.9.12) Line 1, Col 1 ASCII CRLF RW Mem 87%

Type here to search

ENG 11:10 AM
IN 10/15/2022