CUSTOMER CHURN PREDICTION

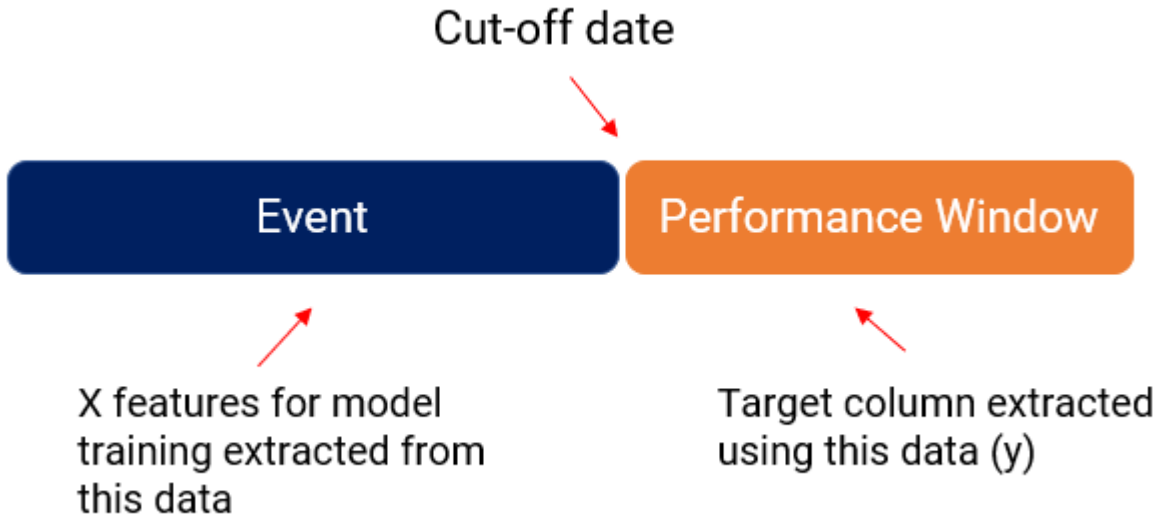
Introduction:

Customer retention is one of the primary KPI for companies with a subscription-based business model. Competition is tough particularly in the SaaS market where customers are free to choose from plenty of providers. One bad experience and customer may just move to the competitor resulting in customer churn.

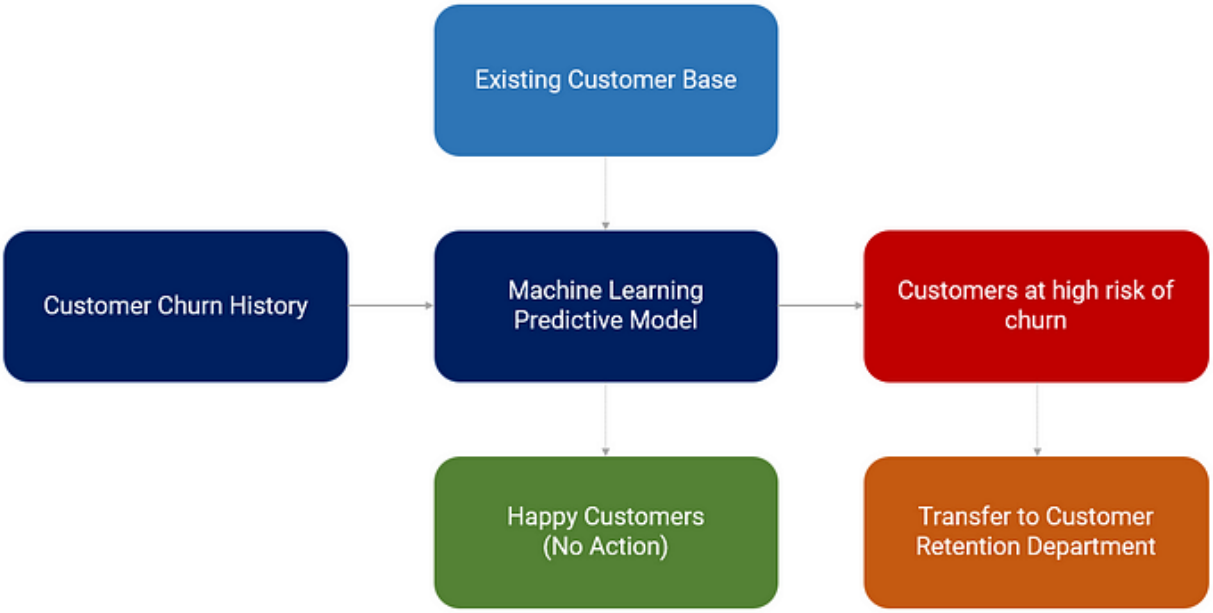
Customer Churn:

Customer churn is the percentage of customers that stopped using your company's product or service during a certain time frame. One of the ways to calculate a churn rate is to divide the number of customers lost during a given time interval by the number of active customers at the beginning of the period. For example, if you got 1000 customers and lost 50 last month, then your monthly churn rate is 5 percent.

There are two broad concepts to understand here:



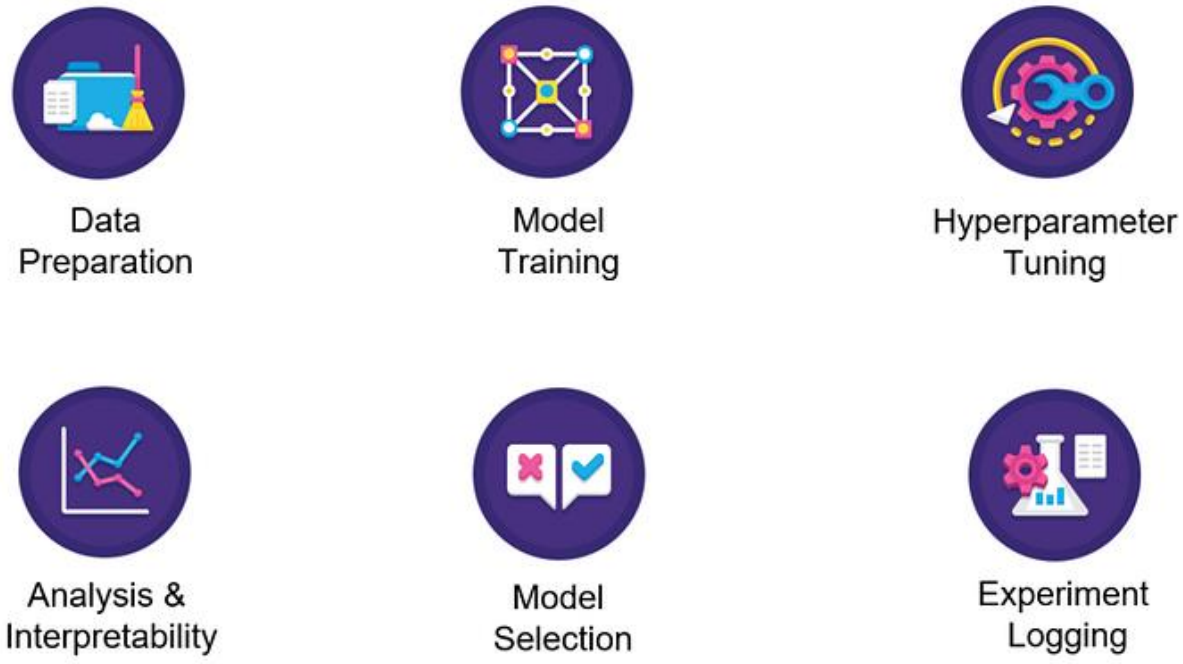
Customer Churn Model Workflow:



Let’s get started with the practical example

PyCaret:

PyCaret is an open-source, low-code machine learning library and end- to-end model management tool built-in Python for automating machine learning workflows. PyCaret is known for its ease of use,simplicity, and ability to quickly and efficiently build and deploy end-to-end machine learning pipelines. To learn more about PyCaret, check out their GitHub .



Features of PyCaret

Install PyCaret

install pycaret

pip install pycaret

Dataset:

```
# import libraries

import pandas as pd

import numpy as np# read csv data

data =

pd.read_csv(&amp;#39;https://raw.githubusercontent.com/srees1988/predict-

churn-py/main/customer_churn_data.csv&amp;#39;)
```

Output:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
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	3668-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	

5 rows × 21 columns

Sample dataset

Exploratory Data Analysis:

```
# check data types

data.dtypes
```

```
customerID           object
gender               object
SeniorCitizen         int64
Partner              object
Dependents           object
tenure                int64
PhoneService         object
MultipleLines        object
InternetService      object
OnlineSecurity       object
OnlineBackup         object
DeviceProtection     object
TechSupport          object
StreamingTV          object
StreamingMovies      object
Contract             object
PaperlessBilling     object
PaymentMethod        object
MonthlyCharges      float64
TotalCharges         object
Churn                object
dtype: object
```

Data types

Model Training & Selection:

Let’s start the training process by using compare_models functionality.This function trains all the algorithms available in the model library and evaluates multiple performance metrics using cross-validation.

compare all models

best_model = compare_models(sort='AUC')

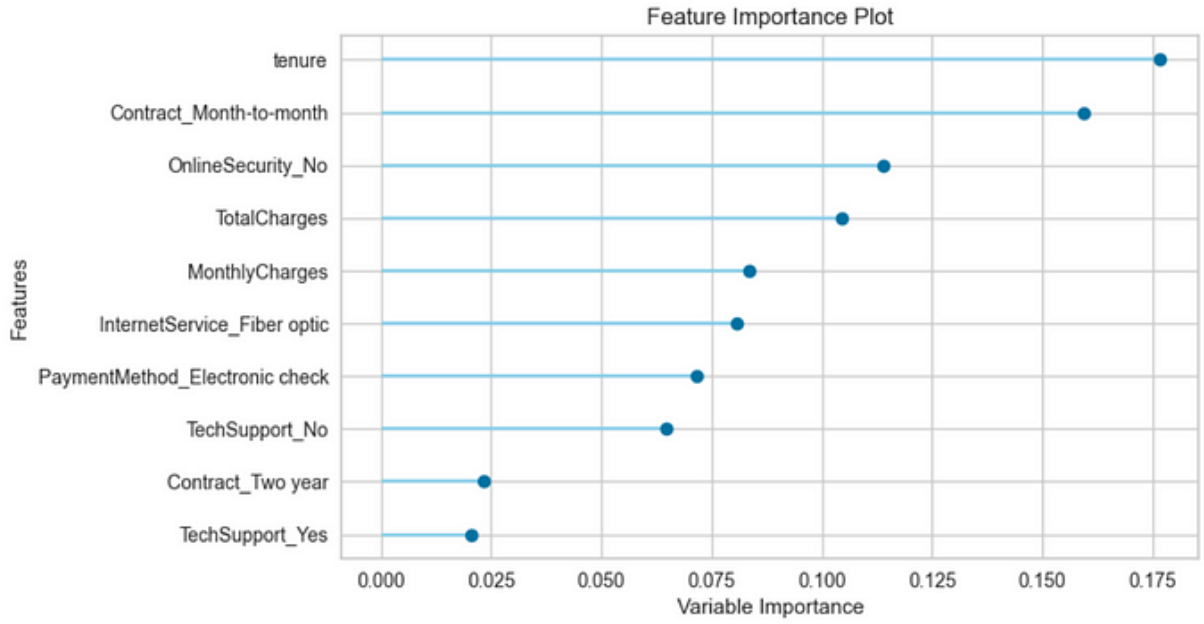
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8002	0.8472	0.5086	0.6475	0.5690	0.4416	0.4474	0.1800
lr	Logistic Regression	0.8061	0.8431	0.5281	0.6594	0.5857	0.4613	0.4666	0.0280
ada	Ada Boost Classifier	0.7992	0.8409	0.5078	0.6449	0.5678	0.4395	0.4450	0.0640
catboost	CatBoost Classifier	0.7943	0.8376	0.5047	0.6312	0.5605	0.4285	0.4333	6.9900
lda	Linear Discriminant Analysis	0.7992	0.8368	0.5366	0.6356	0.5811	0.4505	0.4537	0.0110
lightgbm	Light Gradient Boosting Machine	0.7907	0.8323	0.5148	0.6176	0.5613	0.4254	0.4286	0.0480
nb	Naive Bayes	0.7347	0.8286	0.7777	0.4945	0.6042	0.4195	0.4441	0.0070
xgboost	Extreme Gradient Boosting	0.7878	0.8215	0.4985	0.6141	0.5492	0.4127	0.4171	0.3830
rf	Random Forest Classifier	0.7925	0.8193	0.4844	0.6321	0.5477	0.4163	0.4229	0.2420
et	Extra Trees Classifier	0.7730	0.7881	0.4610	0.5817	0.5133	0.3681	0.3729	0.2550
knn	K Neighbors Classifier	0.7606	0.7492	0.4298	0.5521	0.4827	0.3302	0.3350	0.0740
dt	Decision Tree Classifier	0.7323	0.6565	0.4953	0.4865	0.4903	0.3090	0.3093	0.0100
qda	Quadratic Discriminant Analysis	0.5655	0.6069	0.6929	0.3433	0.4522	0.1631	0.1941	0.0080
svm	SVM - Linear Kernel	0.7347	0.0000	0.4884	0.5776	0.4828	0.3257	0.3543	0.0180
ridge	Ridge Classifier	0.8041	0.0000	0.5039	0.6636	0.5718	0.4480	0.4558	0.0060

The best model based on AUC is Gradient Boosting Classifier . AUC using 10-fold cross-validation is 0.8472.

Model Analysis:

Feature Importance Plot

plot_model(tuned_gbc, plot = 'feature')



Feature Importance Plot