Customer Churn Prediction Report

Description

At Sunbase, we prioritize understanding our customers and ensuring their satisfaction. To achieve this, we want to develop a machine learning model that predicts customer churn.

Objective

Develop a machine learning model to predict customer churn based on historical customer data. Using typical multiple machine learning models, project pipelines, from data preprocessing to model deployment.

Approach

Data Ingestion:-

We'll use **Pandas** library to read and ingest the data and in this process we will also drop the columns that are unnecessary.

EDA:-

We'll perform the EDA over the data using **Pandas-Profilling** library which gives comprehensive report about the data for example:- missing values, correlation and much more.

Categorical and Numerical:-

In this step we'll divide the data according to there data type, if data contains object data type we'll put it in a variable name categorical columns and data other than object type will be put in numerical columns variable. We are dividing data to scale and encode it accordingly.

Pipeline:-

In pipeline we'll use **simple imputer** for imputing the missing values and we'll scale the numerical columns. The categorical columns will be imputed using **simple imputer** and encoded by using **OneHotEncoder**, also **column transformer** will be used to transform the columns.

Train/Test Split:-

After that we'll simply use **train_test_split** available in **Sklearn** for dividing the data in X_train, X_test, y_train, y_test

Model Training:-

Output column churn only contains 0 and 1, In order to predict it we have to use the **Binary Class Classifiers.**

List of the Classifiers We'll train

Logistic Regression
Random Forest Classifier
XGBoost Classifier
CatBoost Classifier
Support Vector Machine (SVM)
Naive Bayes

Model Evaluation:-

For evaluating the model performance over the data we'll use **Accuracy Score** available in Sklearn also to check how well the model is trained we'll use **model.score**.

EDA Observations

1. Here is the sample of given dataset -

Sample

First rows | Last rows

	CustomerID	Name	Age	Gender	Location	Subscription_Length_Months	Monthly_Bill	Total_Usage_GB	Churn
0	1	Customer_1	63	Male	Los Angeles	17	73.36	236	0
1	2	Customer_2	62	Female	New York	1	48.76	172	0
2	3	Customer_3	24	Female	Los Angeles	5	85.47	460	0
3	4	Customer_4	36	Female	Miami	3	97.94	297	1
4	5	Customer_5	46	Female	Miami	19	58.14	266	0
5	6	Customer_6	67	Male	New York	15	82.65	456	1
6	7	Customer_7	30	Female	Chicago	3	73.79	269	0
7	8	Customer_8	67	Female	Miami	1	97.70	396	1
8	9	Customer_9	20	Female	Miami	10	42.45	150	1
9	10	Customer_10	53	Female	Los Angeles	12	64.49	383	1

2. Data has 1 Lakh entries, no missing values, no duplicated entries and It has 9 columns, In which 5 columns are numeric and 4 columns are categorical data types.

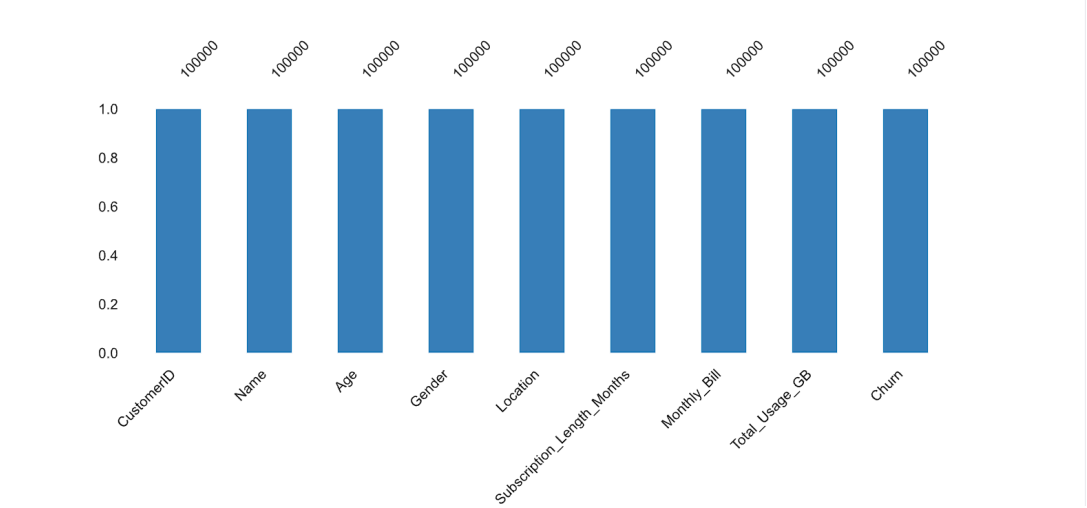
Dataset statistics

Number of variables	9
Number of observations	100000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.9 MiB
Average record size in memory	72.0 B

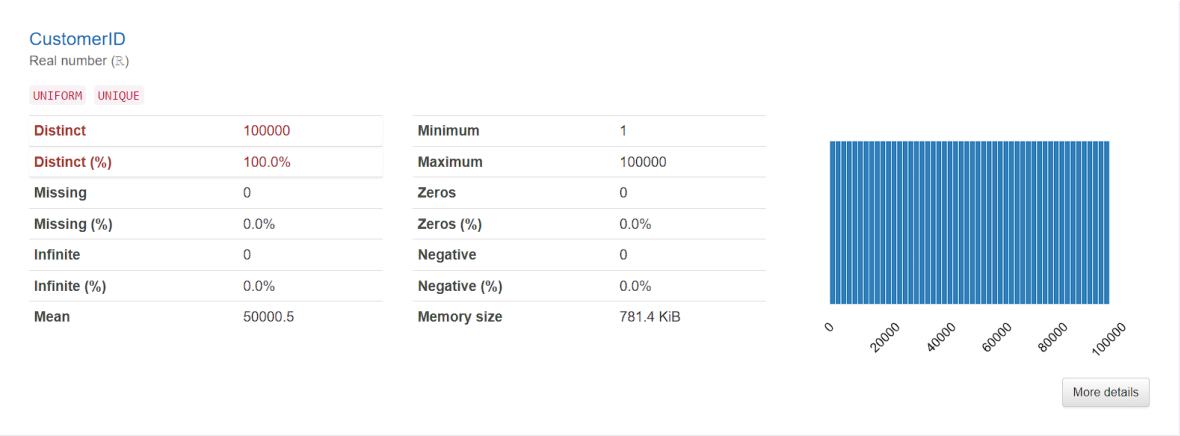
Variable types

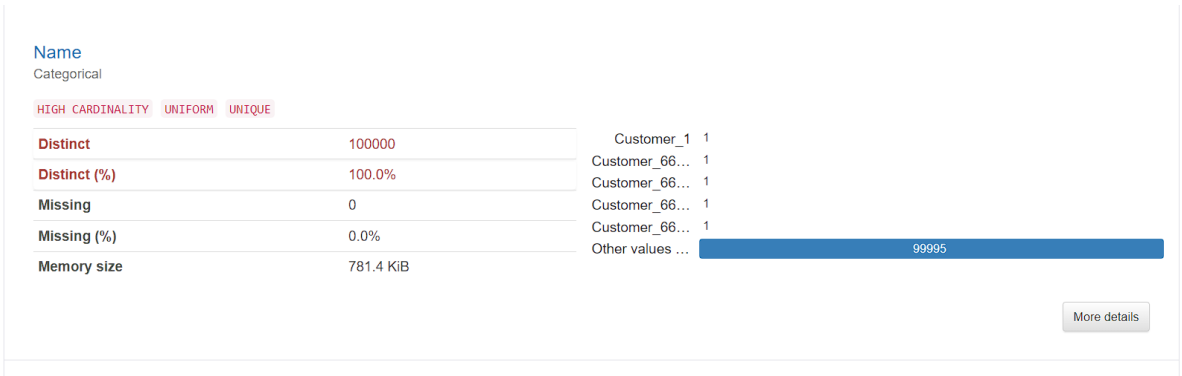
Numeric	5
Categorical	4

3. No missing values in any column here is a count plot



4. Here is the observation of every column

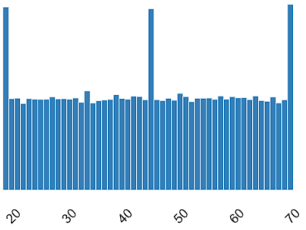




Age

Real number (ℝ)

Distinct	53	**Minimum**	18
Distinct (%)	0.1%	**Maximum**	70
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	44.02702	**Memory size**	781.4 KiB

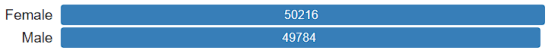


More details

Gender

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	781.4 KiB

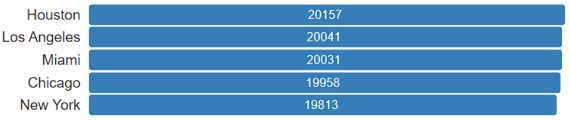


More details

Location

Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	781.4 KiB

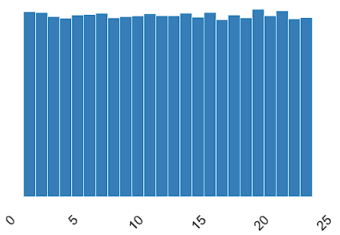


More details

Subscription_Length_Months

Real number (ℝ)

Distinct	24	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	24
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	12.4901	**Memory size**	781.4 KiB

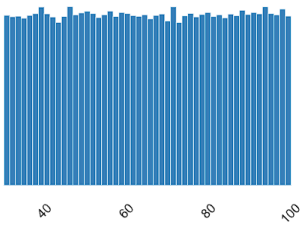


More details

Monthly_Bill

Real number (ℝ)

Distinct	7001	**Minimum**	30
Distinct (%)	7.0%	**Maximum**	100
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	65.053197	**Memory size**	781.4 KiB

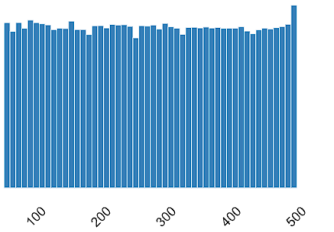


More details

Total_Usage_GB

Real number (ℝ)

Distinct	451	**Minimum**	50
Distinct (%)	0.5%	**Maximum**	500
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	274.39365	**Memory size**	781.4 KiB

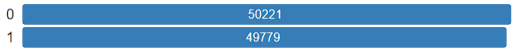


More details

Churn

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	781.4 KiB

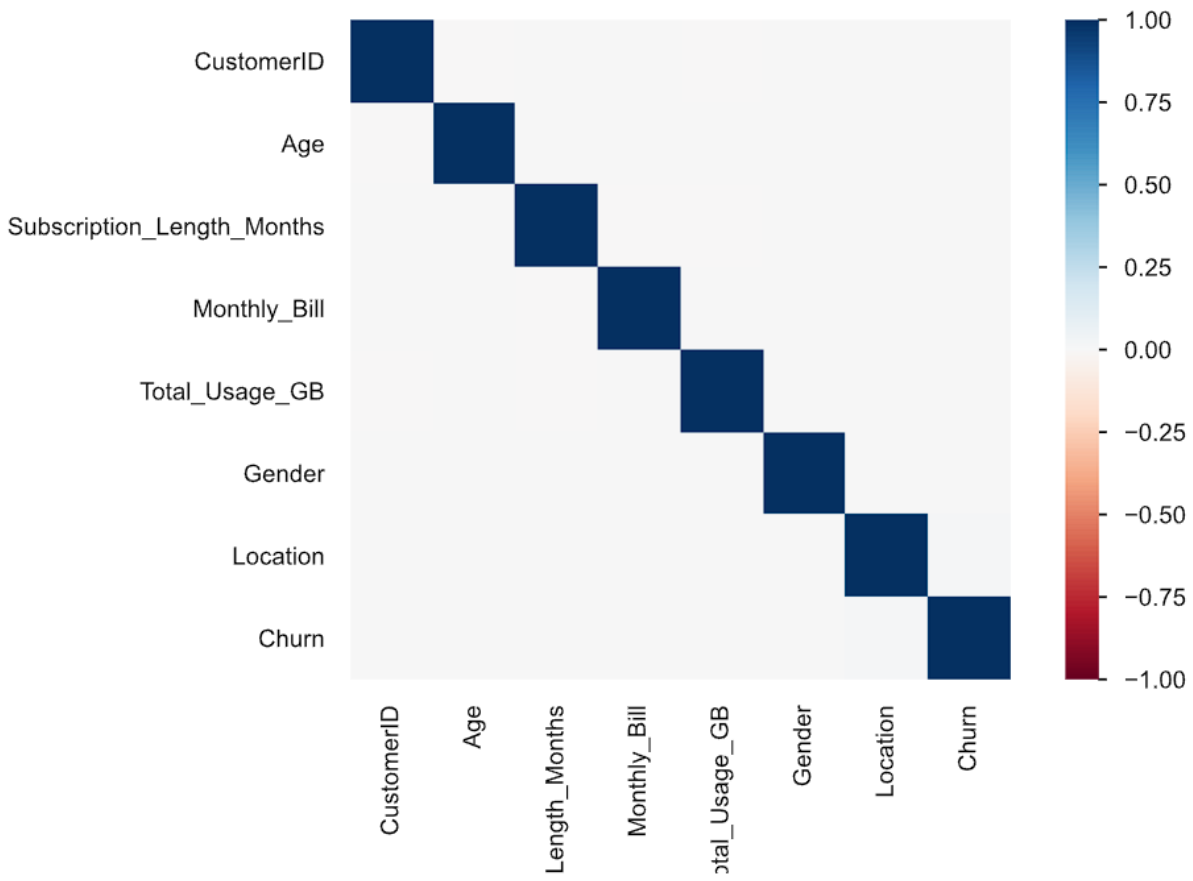


More details

5. Correlation with Heatmap shows that there is correlation but very minute, here's the table and visualisation

This table quantifies the correlation

	CustomerID	Age	Subscription_Length_Months	Monthly_Bill	Total_Usage_GB	Gender	Location	Churn
CustomerID	1.000	-0.001	0.005	0.001	-0.004	0.000	0.007	0.004
Age	-0.001	1.000	0.003	0.001	0.002	0.000	0.000	0.000
Subscription_Length_Months	0.005	0.003	1.000	-0.005	-0.002	0.000	0.000	0.000
Monthly_Bill	0.001	0.001	-0.005	1.000	0.003	0.000	0.002	0.006
Total_Usage_GB	-0.004	0.002	-0.002	0.003	1.000	0.000	0.000	0.006
Gender	0.000	0.000	0.000	0.000	0.000	1.000	0.001	0.000
Location	0.007	0.000	0.000	0.002	0.000	0.001	1.000	0.008
Churn	0.004	0.000	0.000	0.006	0.006	0.000	0.008	1.000



Model Training observation

We have trained all the above mentioned models and there accuracy was somewhat similar to each other but **RandomForestClassifier** gave a better accuracy score and training score, so we are training random forest which is an ensemble technique in Machine Learning.

Observations during model training are:-
1. Models were unable to read the pattern in the data.
2. Accuracy Score is quite low so model with the best accuracy score(RandomForest) is trained.

Hyperparameter Tuning & Cross Validdation

As we all know parameter tuning is an iterative process after alot of tuning we came up with these parameters which were giving us the best accuracy and training score, also we used **GridsearchCv** For cross validation of model and get the best parameters and accuracy score.
Here are the parameters we used-

```
n_estimators=1000,   # number of trees in forest
criterion='gini',    # criterion method
ccp_alpha=0.3,       # cost complexity pruning
max_depth=100,       # maximum depth
min_samples_split=5, # minimum number of samples
min_samples_leaf=2,  # number of leaf
max_features="sqrt", # maximum features
 random_state=42
```

Model Evaluation

We used accuracy score which is available in the sk-learn library, which is a very popular open source library for training score check we used model.score method which returns the accuracy of the trained model on the training data by comparing the predicted labels of the model with actual label.

By Nilay Kumar Sahu