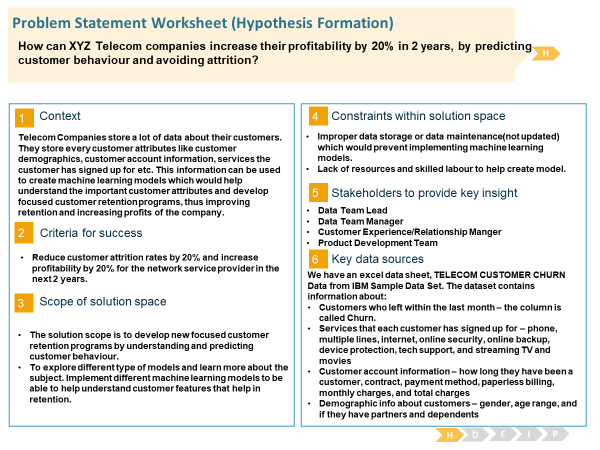
**TELECOM CUSTOMER CHURN ANALYSIS**

**Context:**

Telecom providers find that strong customer retention strategies are becoming increasingly important with the telecommunications industry at a saturation point, and with services largely commodified, providers are suffering from exceptionally high churn rates.

Accenture reports that 77 per cent of consumers are now withdrawing their loyalty faster than they were three years ago and therefore Telco’s must work harder than ever before to retain their customer base. Acquisition costs far outweigh those of maintaining existing customers, further motivating companies to implement innovative strategies to boost the retention of customers in the telecom industry. This is further underlined by Bain & Company research, which suggests that a mere 5% increase in the retention rate of a company can increase profits by 25% to 95%.

This project plans to predict customer behaviour to retain customer and explore differ type of models and learn more about the subject. Implement different machine learning models to be able to help understand customer features that help in retention.

Context: "Predict behaviour to retain customers. You can analyse all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

**DATA:**

The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer, as well as a Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

The dataset is available in:

<https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2017/06/19/guide-to-ibm-cognos-analytics-sample-data-sets>

under Telco churn.

Content:

Each row represents a customer; each column contains customer’s attributes described on the column Metadata.

The data set includes information about:

Customers who left within the last month – the column is called Churn

Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

Customer account information – how long they have been a customer, contract, payment method, paperless billing, monthly charges, and total charges

Demographic info about customers – gender, age range, and if they have partners and dependents

Inspiration: To explore this type of models and learn more about the subject. Implement different machine learning models to be able to help understand customer features that help in retention.

**Demographics**

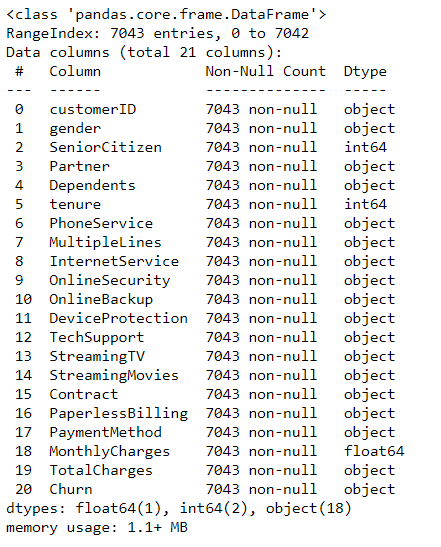
* Gender: The customer’s gender: Male, Female.
* Senior Citizen: Indicates if the customer is 65 or older: Yes, No
* Partner: Indicates if the customer has a partner or not: Yes, No
* Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
* Number of Dependents: Indicates the number of dependents that live with the customer.

**Services**

* Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.
* Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No
* Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, fibre Optic, Cable.
* Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No
* Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No
* Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No
* Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No
* Streaming TV: Indicates if the customer uses their Internet service to stream television programming from a third-party provider: Yes, No. The company does not charge an additional fee for this service.
* Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third-party provider: Yes, No. The company does not charge an additional fee for this service.
* Contract: Indicates the customer’s current contract type: Month-to-Month, One Year, Two Year.
* Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No
* Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check.
* Monthly Charge: Indicates the customer’s current total monthly charge for all their services from the company.
* Total Charges: Indicates the customer’s total charges, calculated to the end of the quarter specified above.

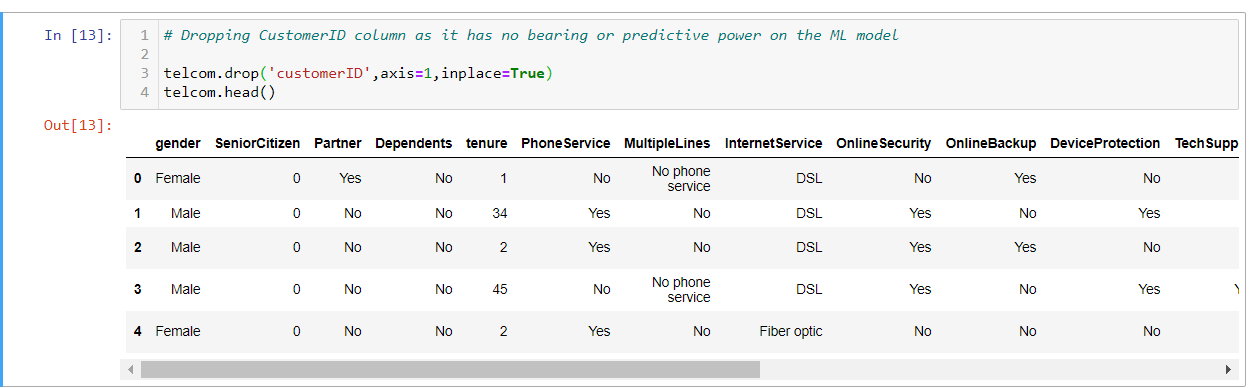
**Status**

* Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

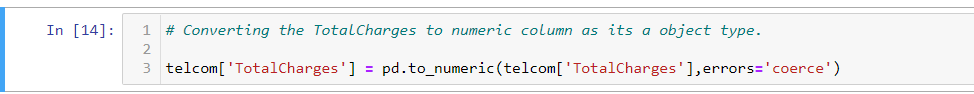


**Data Wrangling:**

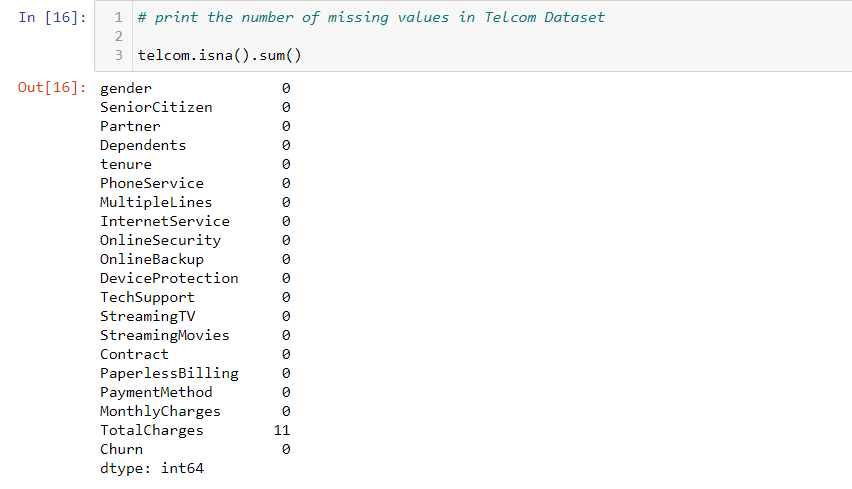
* Dropping CustomerID column as it has no bearing or predictive power on the ML model.



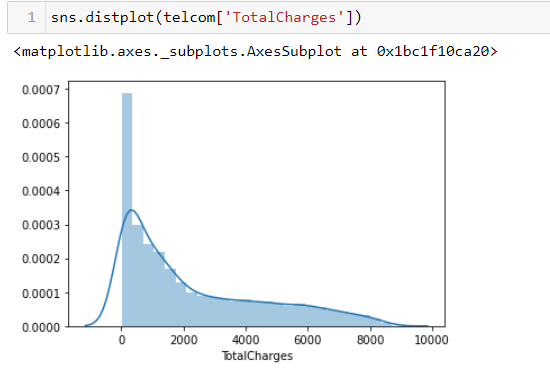
* Converting the TotalCharges to numeric column as it is an object type.



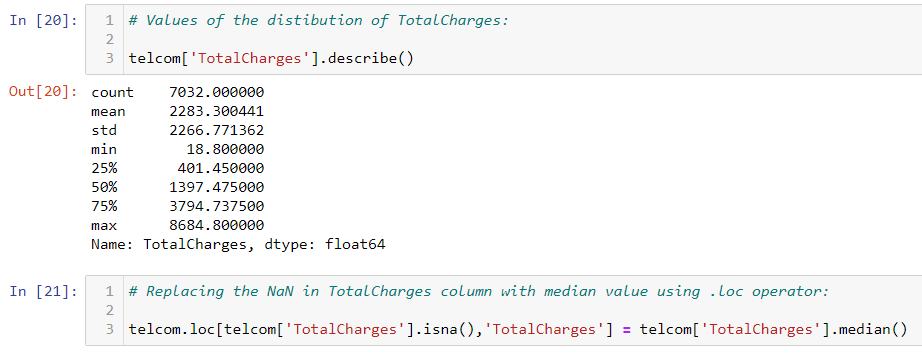
* After converting the Total charges column to numeric we witnessed the presence of 11 null values in the column.



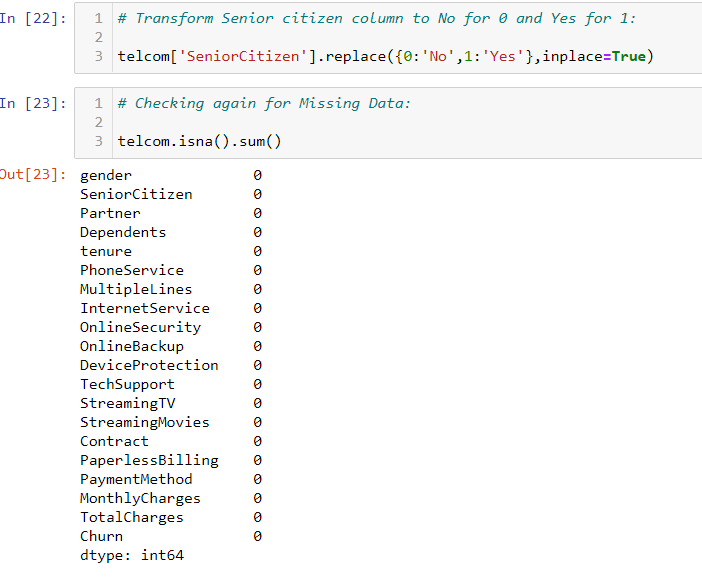
* We look at the distribution of Total Charges column to determine the appropriate value for the column.



* Upon looking at the distribution it would be a better choice to replace the null values with the median value of the column Total charges rather than the mean.



* We also observe that the Senior Citizen column is of integer datatype and replace the values of 0 to ‘No’ and 1 to ‘Yes’, at the same time changing the column data type to object or categorical. We check for missing data again and observe that there are no more missing values in the data frame.



**CLASS IMBALANCE PROBLEM:**

We observe a class imbalance problem for Non Churners(Majority Class) compared to Churners(Minority class).This in the future will have an impact on our models ability to predict correctly the Churners as the model will predominantly be trained to recognise Non Churners correctly but there will be a lot of false positives while predicting the Churners class.

We therefore implemented a resampling technique i.e. Under sampling by deleting random observations from the majority class in this case Non-Churners in order to match the numbers of minority class (Churners).



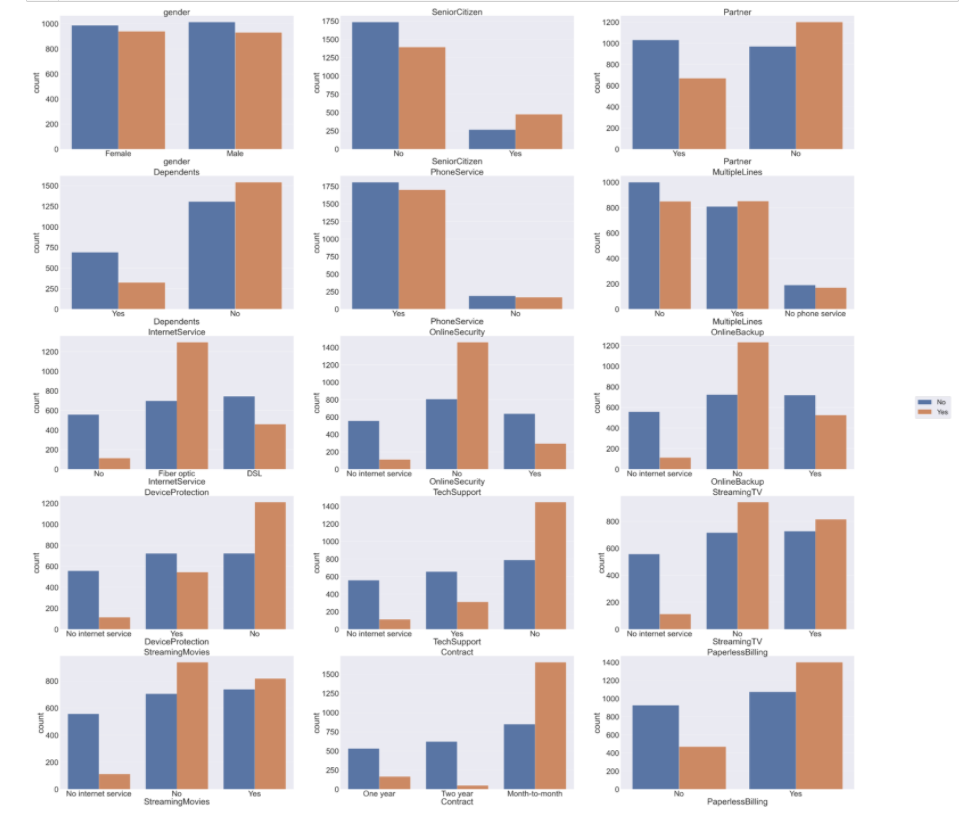
**Exploratory Data Analysis on Balanced Dataset:**

We carried out EDA to understand the relationship between our target/response variable i.e. ‘CHURN’ and other independent variables in the data set.

First segregated the categorical variables from the dataset and plotted a count plot graph to understand the dynamics with the dependent variable ‘CHURN’.

We observed that:

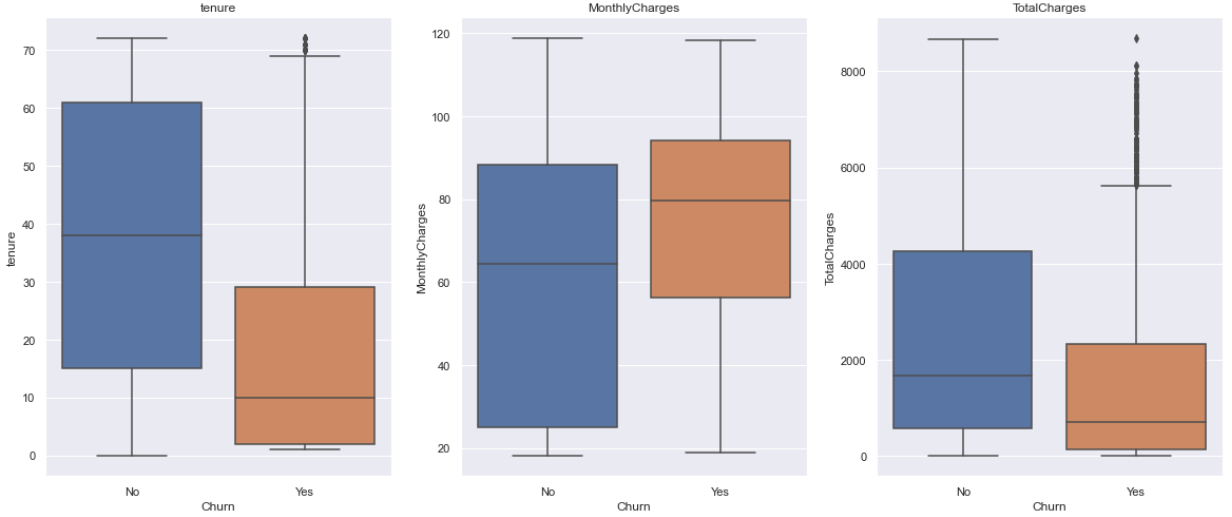
* Observation 1: Gender has 50-50 split on Churn.
* Observation 2: Fiber Optics in Internet Services has a lot of Churners.
* Observation 3: Users not having Online Security, Online Backup, DeviceProtection, TechSupport seem to Churn.
* Observation 4: Users with no partners, dependents tend to churn more.
* Observation 5: Its interesting to note that users who have opted for Streaming Services like StreamingTV, StreamingMovies also have equally high number of Churners.
* Observation 6: Users who are subscribed to monthly billing churn more.



After this we visualise the target variable relationship with the numerical/continuous features in the dataset. We plot a boxplot to help understand this dynamic.

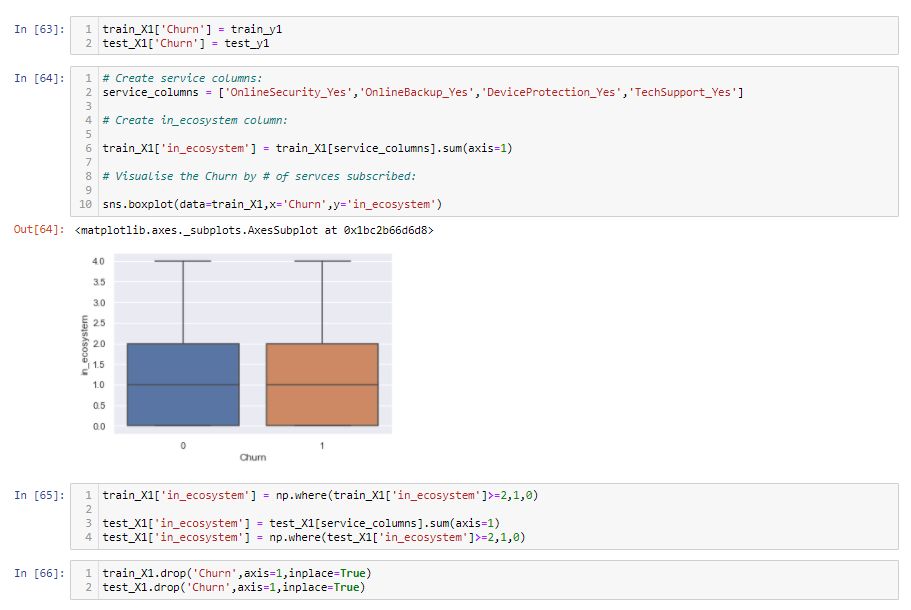
We observed that:

* Observation 1: High Monthly Charges is related to CHURN.
* Observation 2: Churners have relatively low tenure compared to Non-Churners which is understandable.



**Feature Engineering:**

While viewing the Categorical features in EDA for telco dataset we observed that customers who had subscribed to Online Security, Online Backup, DeviceProtection, TechSupport were less prone to Churn. We add a new feature called **in\_ecosystem** which lets us know the count of services a customer has subscribed to.

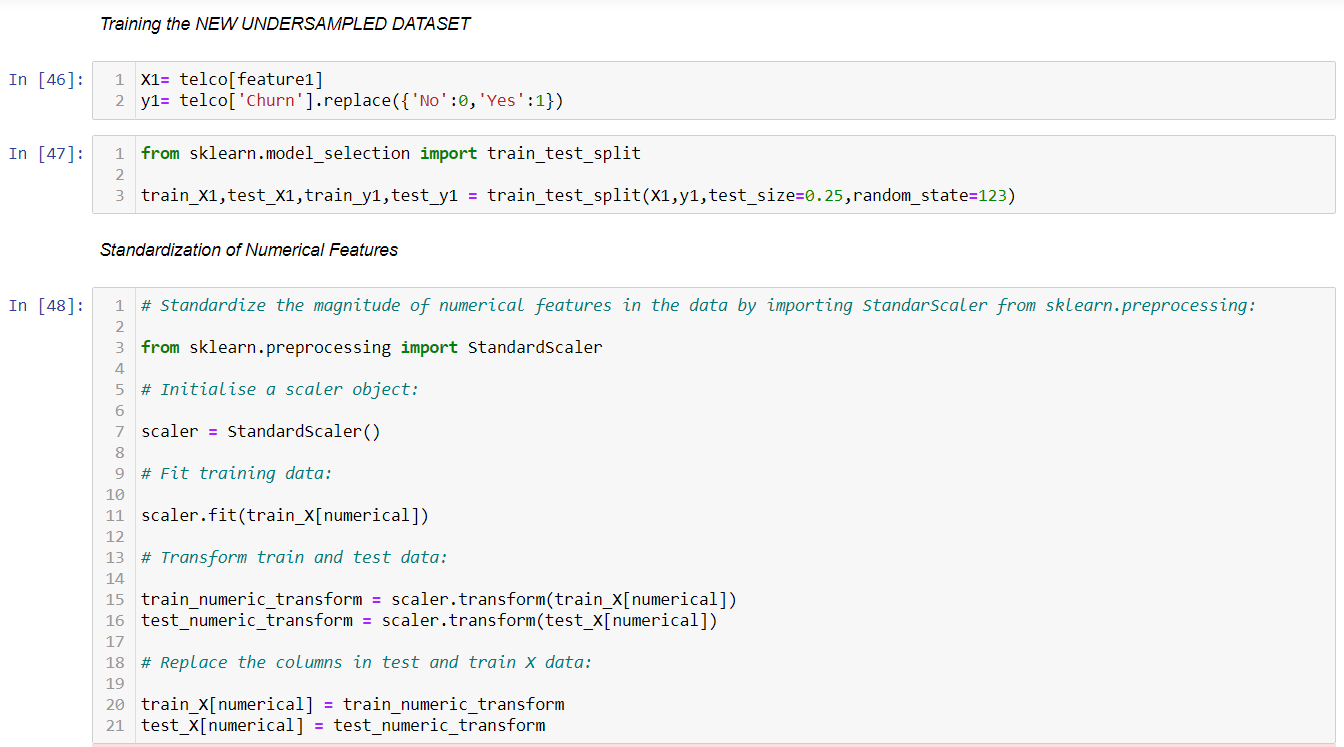


**Data Pre-processing and Modelling:**

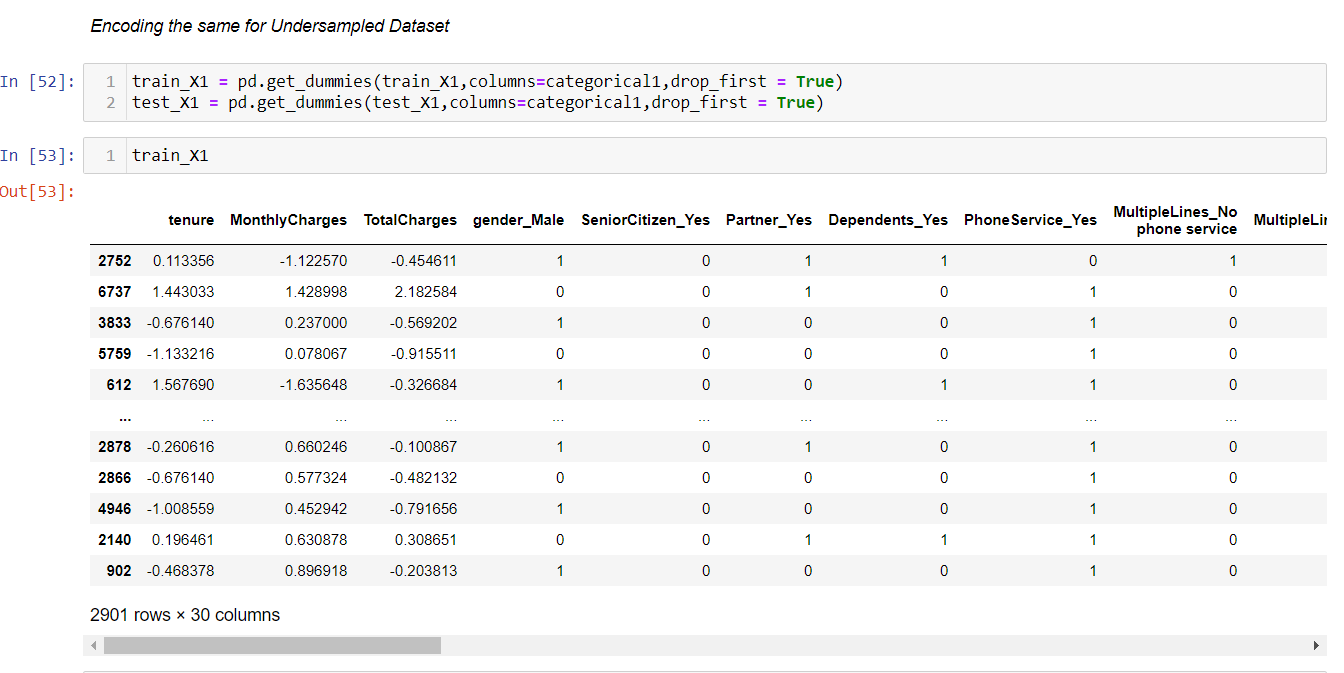
Data pre-processing was carried out to account for the differences in magnitude of numeric features in the dataset (Standardizing) and for situations where one has categorical and continuous variables.

The following steps were carried out:

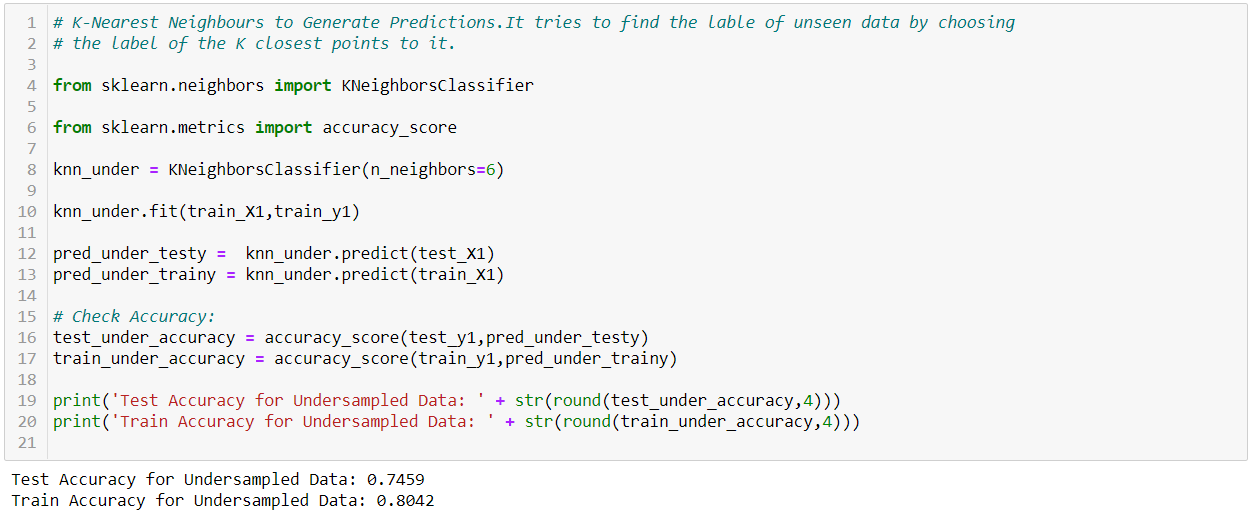
1. Split the original dataset into Training and Test Datasets.
2. Standardise the magnitude of numeric features in the data frame.



1. Create dummy variable or indicators features for categorical variables in the dataset.



**MODELLING:** I tried out with the simplest model K-Nearest Neighbours to predict customer churn. K-Nearest Neighbours tries to find the label of unseen data by choosing the label of the K closest points to it. Below are the results:

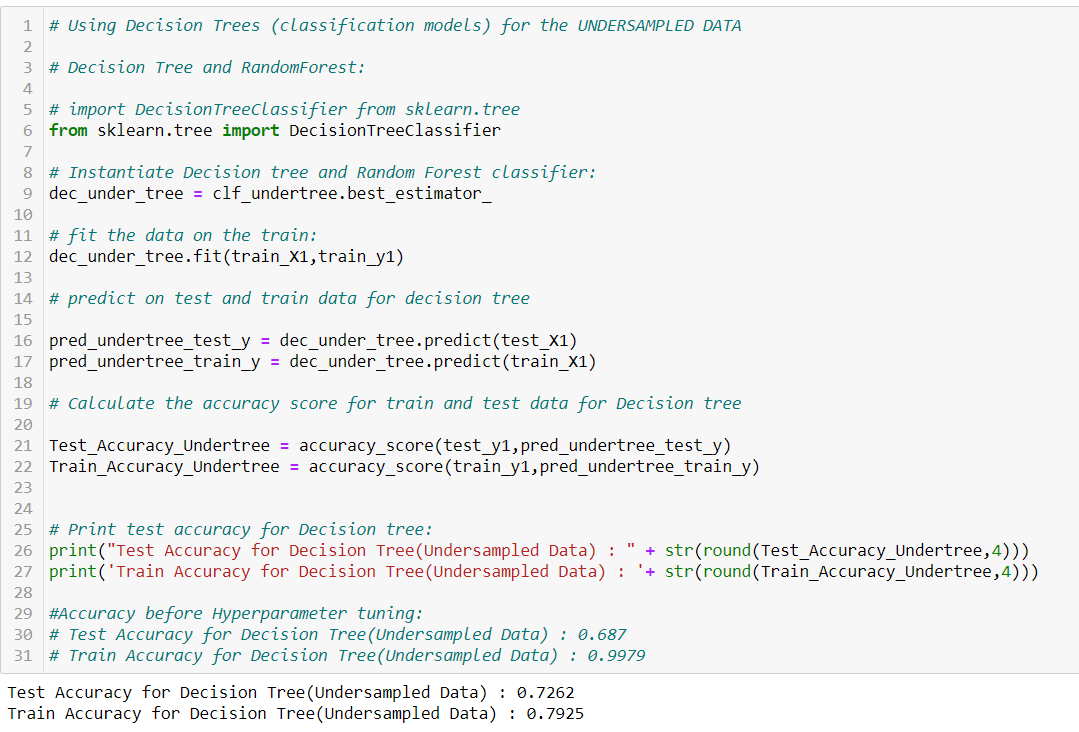


**KNeighborsClassifier Model Accuracy:**

Achieved an accuracy score of 0.7459 on the TEST data and of 0.8042 on the TRAIN data. Since this a preliminary model study I explored other models which could return a higher accuracy score and better confidence results.

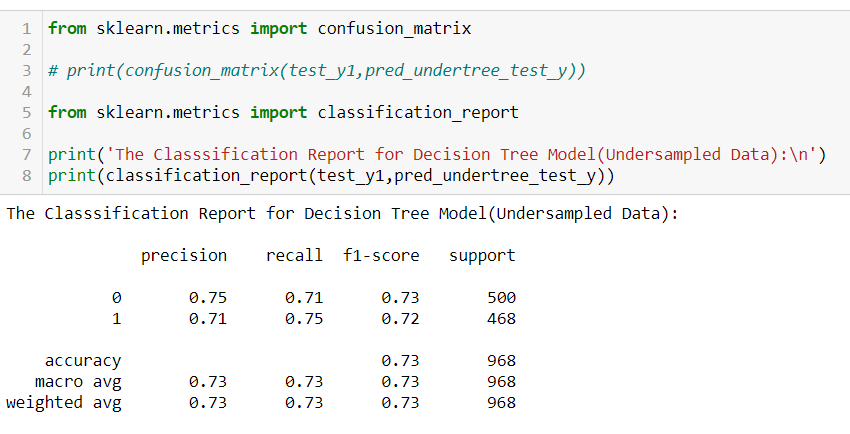
**Decision Tree:**

I implemented Decision Tree model which is a recursive algorithm that ask if-else questions on the data using a set of cut off points. Its aim is to maximise purity(homogeneity) of the resulting data points. Below are the results:



**DecisionTreeClassifier model:**

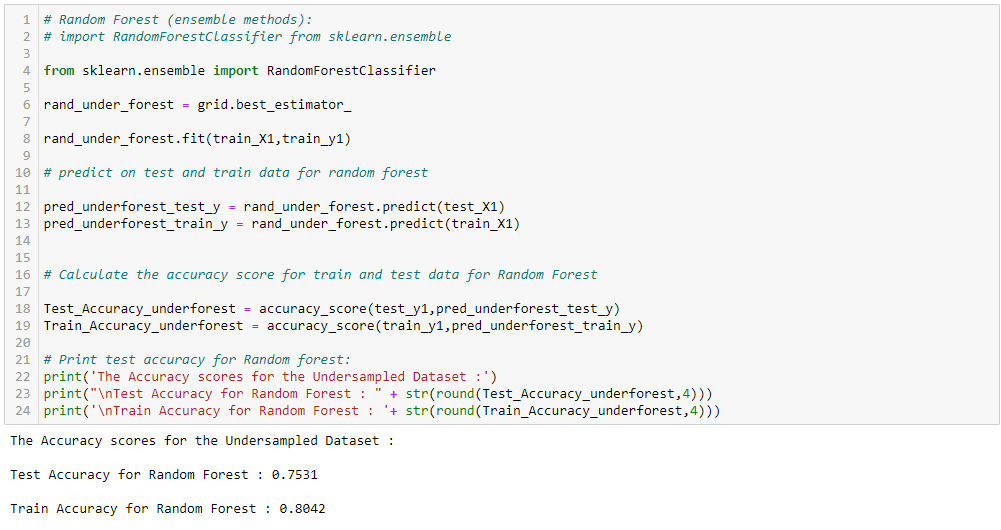
TEST Accuracy Score: 0.7262 and TRAIN Accuracy Score: 0.7925.



Above is the classification report for the Decision Tree model.

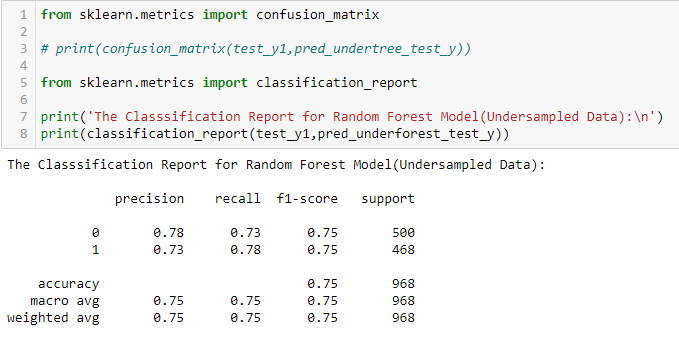
Random Forest:

Random Forest pools the predictions of many decision trees together, fitting data points on a random number of features and samples and returns the most common class. It is an upgrade to Decision Tree model.



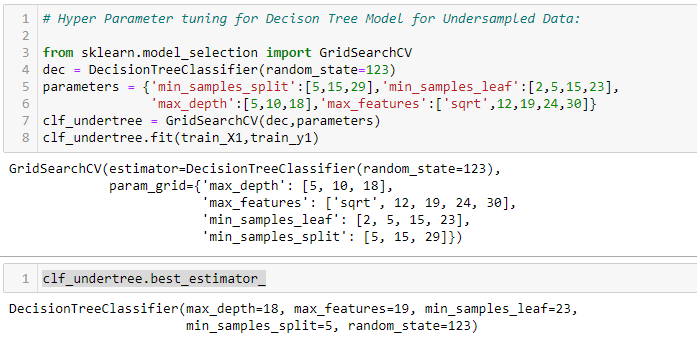
RandomForestClassifier Model Results:

**TEST accuracy score 0.7531 and TRAIN accuracy score 0.8042**. Which is among the best results observed till now for all the models train and test. The classification report for Random Forest model is as follows:



**MODEL HYPERPARAMETER TUNING:**

In the above two models (Decision Tree and Random Forest) I have implemented Hyperparameter tuning. While implementing the models I noticed that the simplest model was not a good enough solution and complex model performance was not meeting requirements, hence, to improve the model accuracy I implemented GridSearchCV for model parameter optimisation.

Decision Tree Hyperparameter tuning:

Random Forest Hyperparameter tuning:

