**Problem Statement explanation:**

The telecommunications industry experiences a **churn of 15 – 25%** a year and it costs on average **5-10 times** more to acquire a new customer rather than to retain an existing one. Customer retention has now become more important than customer acquisition. To reduce customer churn, telecom companies need to predict which customers are at a **high risk of churn**.

Business Objective:

1. Building predictive models to identify customers that are at a high risk of churn
2. Identify the main indicators of churn

Ways to define churn:

* **Usage-based churn**: Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period.
* Revenue-Based churn: Customers who have not utilized any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period.

**80-20 rule**: 80% of revenue comes from the top 20% customers (called high-value customers)

The three phases of customer churn: The **good** phase, the **action** phase and the **bad** phase

The idea will be to filter out the HVC or high-value-customers on the basis of first and second month, derive features to understand overall usage, tag the customers among this who have churned using the following metrics:

1. Total incoming calls
2. Total outgoing calls
3. 2g data usage
4. 3g data usage

Then, we remove the attributes that correspond to the month 9 of the bad phase

**Data Interpretation, Preparation and EDA**:

1. Import the required functions, do a **sense check** of the data, look at the **% of missing values** (99999 rows and 226 columns)
2. **Impute** the missing values using median, mode. **Remove** the columns that have **constant variance** or the same value throughout. Impute the meaningful NA values with 0 and re-calculate % of missing values (done on the entire dataset). Drop the columns that have over **70% of missing values** post imputation, drop **identifiers**
3. Iterative Imputer used to impute values between 1 – 8%, the estimator used was Bayesian Ridge – Round-robin fashion
4. From the date columns, get a few derived metrics such as the day of the week, the number of days since last recharge etc. – remove the date columns. The data has no categorical values to handle or perform one-hot encoding on
5. Calculate derived columns –
   1. total calls = incoming + outgoing for months 6,7
   2. total data = 3g data usage + 4g data usage
   3. average usage = (total calls + total data) / 2
6. Obtain top 30 percentile (80-20 rule) -> The number is 762 and from 100000 rows, we are left with around 30,000 customer level data

Condition for churn => if total calls + total data usage = 0, then, churn = 1, else churn = 0 (for month 9)

Churned customers for the top 30000 rows is now obtained. We see that there is an imbalance (7.6% of customers have churned)

1 = 2284 || 0 = 27737

We now proceed to perform EDA:

* We have already performed missing value analysis
* We perform Outlier analysis
* We remove the column that have low variance
* Remove the flags that we made, remove the columns that have constant value

We split into train-test and then proceed to perform feature scaling individually on the train data and then the test data and we handle class imbalance via SMOTE (Synthetic Minority Over-sampling Technique)

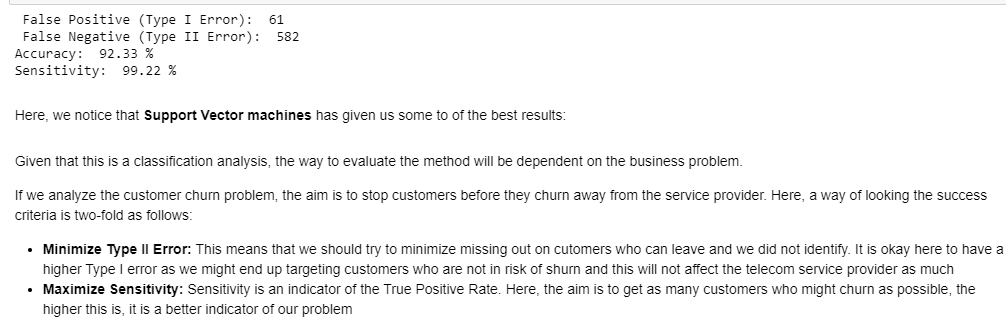
**Model Building and Evaluation**:

Two ways to go about modelling:

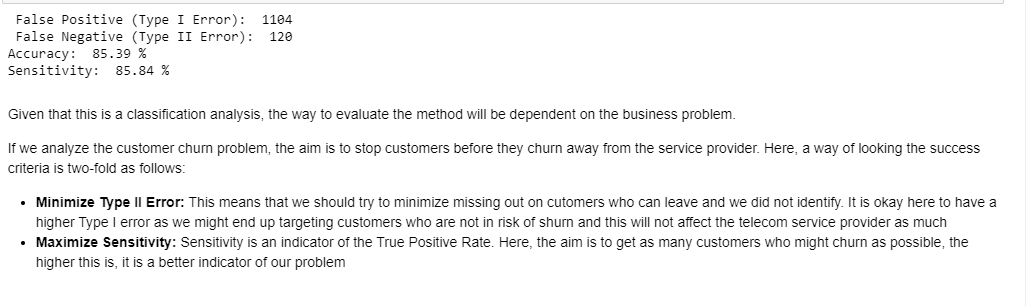
Model 1: PCA + Logistic Regression

1. PCA + LogReg, PCA + Random Forest, PCA + SVM
2. LogReg + +RFE + VIF

SVM with hyperparameter tuning



LogReg + RFE + VIF



Type 2 error (number of False positives) – number of customers who we decide to give offers to if they are not going to churn

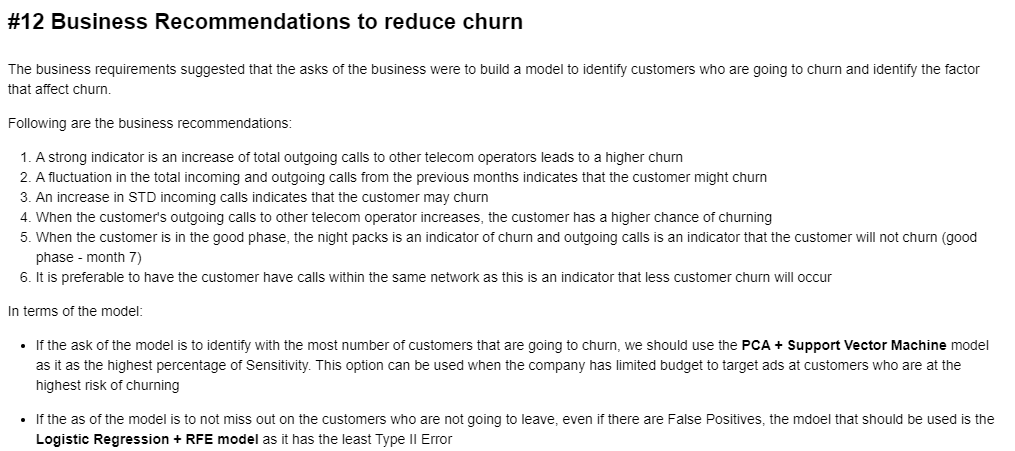
Sensitivity = TP / (TP + FP)

**Financial Benefit Analysis**:

Since it costs 5-10 times to acquire a customer, rather than to retain a customer, it is critical to retain customers. With this aim to retain more customers rather than spend the effort to acquire more, it is critical that this effort is put forth to retain the customers that can add more to the bottom-line of the company rather than to spend money to retain customers who are not high value.

Let’s say the cost of retaining a customer is Rs. 1k and we are successful in preventing 1000 customers from churning. 1,000,000 – we spend 10 lakhs or a million dollars to prevent HVC from churning. Rather, if we were to spend the same time to acquire 1000 customers, per the industry standard, it would cost us 10k to acquire a new customer and to acquire 1000 customers, it would cost us 10 million or a 1 crore to get just as much business.

Hence, running the model at say a monthly or a quarterly cadence, the company would end up spending about 10 lakhs or 0.85 million to retain the customers. If the model was not run, then the company would end up spending 10 million, thus saving the company 9.2 million for every 1000 customers it is able to retain, which has tremendous financial value.



**Subject Matter Understanding**:

**Soft Skills**: