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 –
    a. total calls = incoming + outgoing for months 6,7
    b. total data = 3g data usage + 4g data usage
    c. 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)

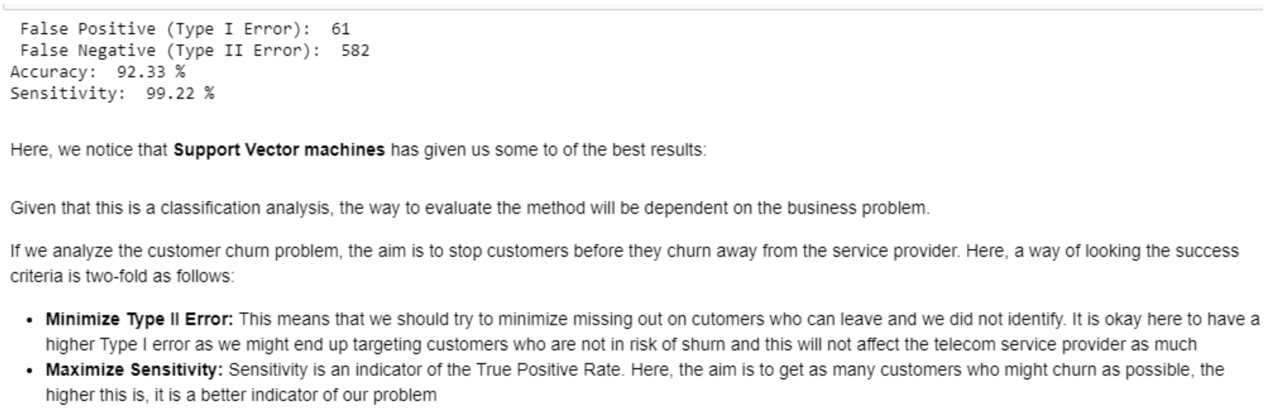
Model Building and Evaluation:

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



```
 False Positive (Type I Error):  61
 False Negative (Type II Error):  582
Accuracy:  92.33 %
Sensitivity:  99.22 %
```

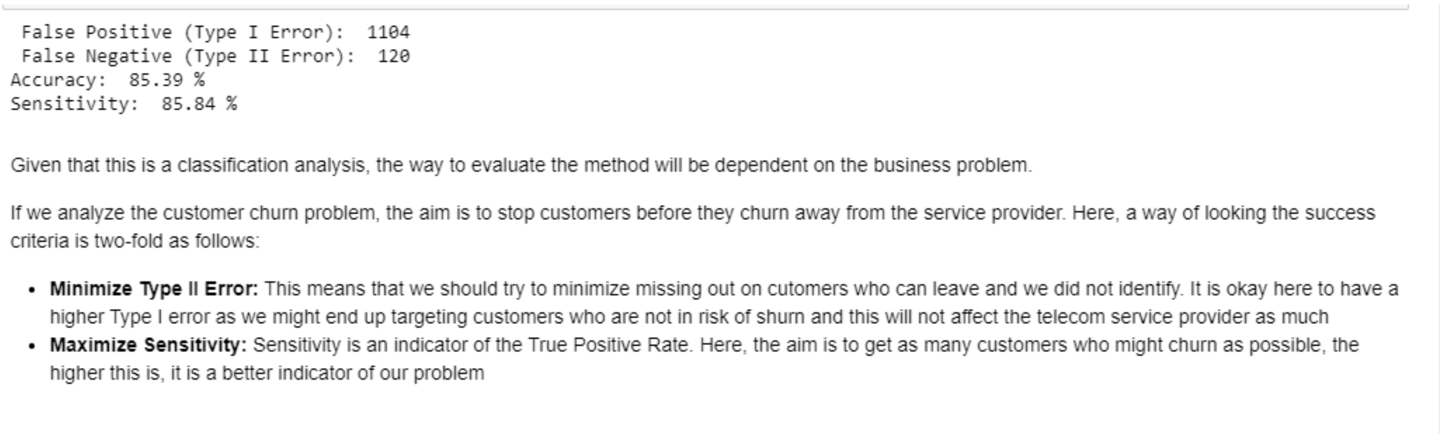
Here, we notice that **Support Vector machines** has given us some to of the best results:

Given that this is a classification analysis, the way to evaluate the method will be dependent on the business problem.

If we analyze the customer churn problem, the aim is to stop customers before they churn away from the service provider. Here, a way of looking the success criteria is two-fold as follows:

- **Minimize Type II Error:** This means that we should try to minimize missing out on cutomers who can leave and we did not identify. It is okay here to have a higher Type I error as we might end up targeting customers who are not in risk of shurn and this will not affect the telecom service provider as much
- **Maximize Sensitivity:** Sensitivity is an indicator of the True Positive Rate. Here, the aim is to get as many customers who might churn as possible, the higher this is, it is a better indicator of our problem

LogReg + RFE + VIF



```
 False Positive (Type I Error):  1104
 False Negative (Type II Error):  120
Accuracy:  85.39 %
Sensitivity:  85.84 %
```

Given that this is a classification analysis, the way to evaluate the method will be dependent on the business problem.

If we analyze the customer churn problem, the aim is to stop customers before they churn away from the service provider. Here, a way of looking the success criteria is two-fold as follows:

- **Minimize Type II Error:** This means that we should try to minimize missing out on cutomers who can leave and we did not identify. It is okay here to have a higher Type I error as we might end up targeting customers who are not in risk of shurn and this will not affect the telecom service provider as much
- **Maximize Sensitivity:** Sensitivity is an indicator of the True Positive Rate. Here, the aim is to get as many customers who might churn as possible, the higher this is, it is a better indicator of our problem

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

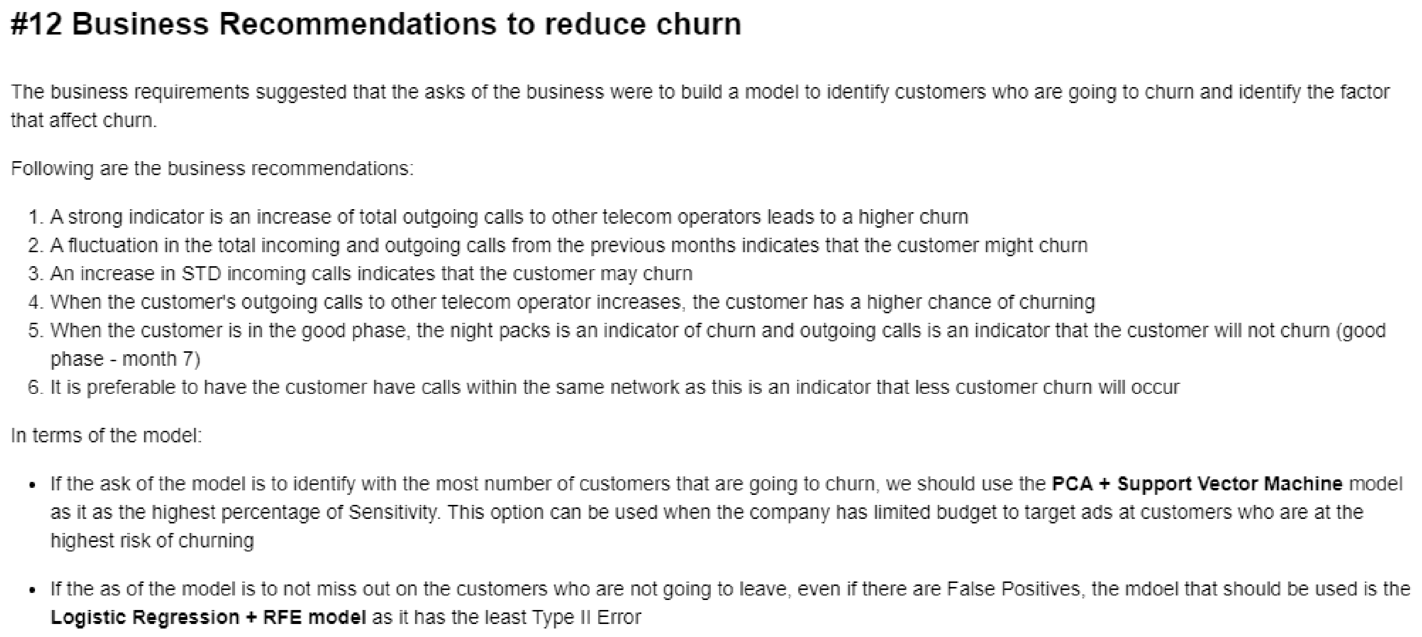
Sensitivity = TP / (TP + FP)

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.



#12 Business Recommendations to reduce churn

The business requirements suggested that the asks of the business were to build a model to identify customers who are going to churn and identify the factor that affect churn.

Following are the business recommendations:

1. A strong indicator is an increase of total outgoing calls to other telecom operators leads to a higher churn
2. A fluctuation in the total incoming and outgoing calls from the previous months indicates that the customer might churn
3. An increase in STD incoming calls indicates that the customer may churn
4. When the customer's outgoing calls to other telecom operator increases, the customer has a higher chance of churning
5. When the customer is in the good phase, the night packs is an indicator of churn and outgoing calls is an indicator that the customer will not churn (good phase - month 7)
6. It is preferable to have the customer have calls within the same network as this is an indicator that less customer churn will occur

In terms of the model:

- If the ask of the model is to identify with the most number of customers that are going to churn, we should use the **PCA + Support Vector Machine** model as it as the highest percentage of Sensitivity. This option can be used when the company has limited budget to target ads at customers who are at the highest risk of churning

- If the as of the model is to not miss out on the customers who are not going to leave, even if there are False Positives, the mdoel that should be used is the **Logistic Regression + RFE model** as it has the least Type II Error

Subject Matter Understanding:

Soft Skills: