# Churn Prediction and Prevention System for Subscription Services Using Machine Learning

- Angadi Abhinay

## 1. Problem Statement:

In the competitive landscape of subscription-based services, reducing customer churn (cancellation of subscriptions) is crucial for business sustainability and growth. The project aims to develop a sophisticated Churn Prediction and Prevention System that utilizes data analytics and machine learning techniques to forecast potential churn and suggest proactive strategies to retain subscribers.

**What is Customer Churning:**

- Customer churn, also known as customer attrition or customer turnover, refers to the phenomenon where customers or clients cease doing business with a company or stop using its products or services. In simpler terms, it's the rate at which customers stop being customers. Churn can occur for various reasons, such as dissatisfaction with the product or service, better alternatives from competitors, changes in customer needs, or economic factors.
- Churn is a critical metric for businesses to monitor, especially in subscription-based models or industries where customer retention is crucial for long-term success. High churn rates can indicate underlying problems with a company's offerings, customer service, or overall customer experience. On the other hand, low churn rates suggest that customers are satisfied and loyal to the company.
- Reducing churn and retaining customers often involve strategies like improving the quality of products or services, enhancing customer support, offering loyalty programs, engaging customers through personalized marketing, and addressing any issues that might be leading to dissatisfaction.

## 2. Market/Customer/Business Need Assessment:

- In comprehensive evaluation of the market conditions, customer requirements, and business needs that justify the development and implementation of a churn prediction and prevention system using machine learning techniques. This assessment aims to understand the reasons for churn in subscription services and how addressing these issues can provide value to the business and its customers.

**Market Need Assessment:**

- What is the overall churn rate in the subscription services industry?
- Are there any emerging trends or challenges in customer retention within this market?
- Are there existing solutions or competitors addressing churn in similar ways?
- What potential benefits can a churn prediction and prevention system offer to the market?

**Customer Need Assessment:**

- What are the primary reasons customers cancel their subscriptions?
- What factors contribute to customer satisfaction and loyalty?
- How can a churn prediction and prevention system enhance the customer experience?
- What features or approaches would customers value in such a system?

## 3. Target Specifications and Characterization:

- Churn Prediction Model: The development of accurate machine learning models that can predict which customers are likely to churn (cancel their subscription) in the near future.
- Churn Prevention Strategies: Identification and implementation of strategies to prevent or mitigate customer churn based on the predictions made by the model. This could involve personalized offers, discounts, or targeted communications to retain customers.
- Model Performance Metrics: Setting clear criteria for evaluating the performance of the churn prediction model. This might include metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve.
- Integration with Existing Systems: If the subscription service already has customer data and systems in place, the target specifications might involve integrating the developed churn prediction and prevention system seamlessly into the existing infrastructure.

## 4. External Search (online information sources/references/links):

we use a Telecom Churn Prediction dataset from Kaggle, which is quite popular for churn modeling. Each row represents a customer, and each column contains the customer's attributes

The dataset contains information about:

- Customers who left – the column is called "Churn", and this will be the target column in our Model (something we want to predict)
- 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've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age, and if they have partners and dependent.

***Sources:***

- Hrvoje Smolic, born in 1976 in Zagreb, Croatia, is the accomplished Founder and CEO of Graphite Note. He holds a Master's degree in Physics from the University of Zagreb. In 2010 Hrvoje founded Qualia, a company that created BusinessQ, an innovative SaaS data visualization software utilized by over 15,000 companies worldwide.

***Relevant Papers:***

- Relevant information about the dataset and corresponding details about the project is explained in the below web page link .

{

## 5. Benchmarking alternate Products:

- Benchmarking involves evaluating the proposed methods against existing approaches or alternate models to determine which one performs better in terms of accuracy, precision, recall, and other relevant metrics.
- Here's how benchmarking alternate products might be applied to our project:

- Selection of Alternate Models
- Data Preparation
- Feature Engineering
- Model Training and Validation
- Performance Metrics
- Comparison and Analysis
- Fine-Tuning
- Implementation and Monitoring

## Our Data set:

```
In [1]: import pandas as pd
        import numpy as np
```

```
In [2]: df = pd.read_csv('/kaggle/input/telco-customer-churn/WA_Fn-UseC_-Telco-Customer-Churn.csv')
        df
```

Out[2]:

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | ... | DeviceProtection | TechS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 7038 | 6840-RESVB | Male | 0 | Yes | Yes | 24 | Yes | Yes | DSL | Yes | ... | Yes | |
| 7039 | 2234-XADUH | Female | 0 | Yes | Yes | 72 | Yes | Yes | Fiber optic | No | ... | Yes | |
| 7040 | 4801-JZAZL | Female | 0 | Yes | Yes | 11 | No | No phone service | DSL | Yes | ... | No | |

## Info about Data set:

```
In [6]: df.info()
```

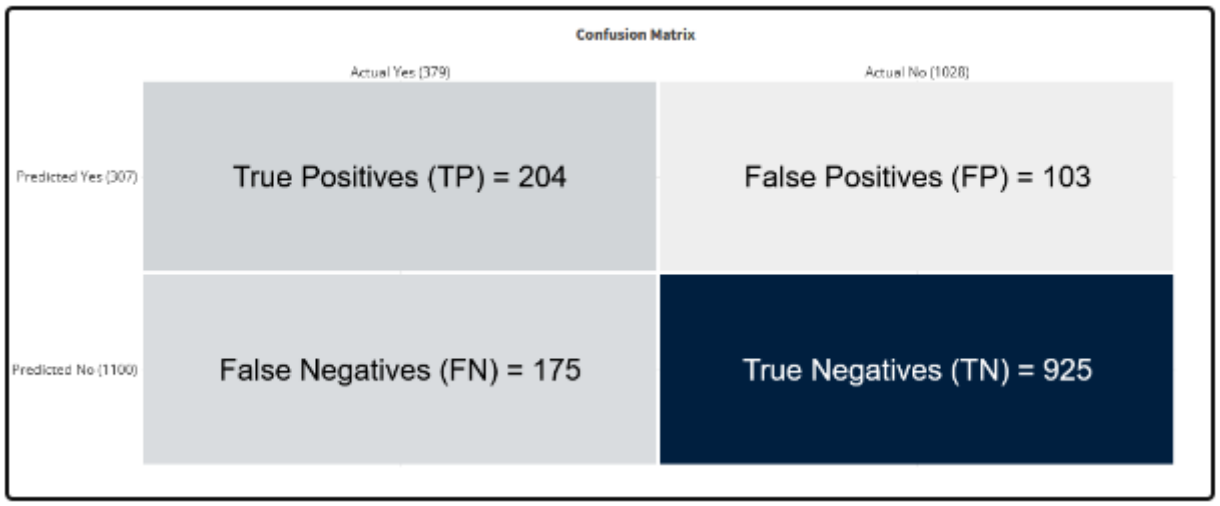
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
```

## Customer Churn Analysis:

- Null values handling
- missing values
- One Hot Encoding
- fix imbalance
- normalization
- constants
- cardinality
- will take a sample of 80% (5625 rows) of our data and train several machine learning models.
- Then, it will test those models on the remaining 20% (1407 rows) and calculate relevant model scores. Based on scores, it will select the best performing model for the dataset.
- The best model fit, results, and predictions are available on the Results tab, after about 20 seconds training.
- In our case the best Model based on the F1 value score is Logistic Regression. Other models' training metrics are listed below.

**Confusion Matrix**:

- Confusion Matrix makes it easy to see whether the Model is confusing two classes (YES and NO in our case). For each class, it summarizes the number of correct and incorrect predictions. The Model predicted column 'Churn' for a test dataset of 1407 rows and compared the predicted outcomes to the historical outcomes.




**Correct Predictions:**

- 1129 in total out of 1407 test rows. This is defining Model Accuracy = 80.24%
- True Positives (TP) = 204: a row was Yes and the model predicted a Yes class for it.
- True Negatives (TN) = 925: a row was No and the model predicted a No class for it.

**Errors:**

- 278 in total out of 1407 test rows, 19.76%
- False Positives (FP) = 103: a row was No and the model predicted a Yes class for it.
- False Negatives (FN) = 175: a row was Yes and the model predicted a No class for it.

**Other Model Scores:**

- Accuracy, (TP + TN) / TOTAL.
- From all the classes (positive and negative), 80.24% of them we have predicted correctly.
- Accuracy should be as high as possible.
- Precision, TP / (TP + FP).
- From all the classes we have predicted as positive, 66.45% are actually positive.

- Precision should be as high as possible.
- Recall, TP / (TP + FN).
- From all the positive classes, 53.83% we predicted correctly.
- Recall should be as high as possible.
- F1 score, 2 * (Precision * Recall)/(Precision + Recall).
- F1-score is 59.48%. It helps to measure Recall and Precision simultaneously.

# 6. Applicable Regulations:

The patents mentioned above might claim the technology used if the algorithms are not developed and optimised individually and for our requirements. Using a pre-existing model is off the table if it incurs a patent claim.

1. Must provide access to the 3rd party websites to audit and monitor the authenticity and behavior of the service.
2. Enabling open-source, academic and research community to audit the Algorithms and research on the efficacy of the product.
3. Laws controlling data collection : Some websites might have a policy against collecting customer data in form of reviews and ratings.
4. Must be responsible with the scraped data : It is quintessential to protect the privacy and intention with which the data was extracted.

# 7. Applicable Constraints:

- Continuous data collection and maintenance
- Lack of technical knowledge for the user
- Taking care of rarely bought products

# 8. Business Model:

**Subscription Licensing:**:

- Offer the churn prediction and prevention system as a subscription-based service to other businesses that provide subscription services. Charge these businesses based on the size of their customer base or the frequency of predictions made.

**Pay-Per-Use:**

- Implement a pay-per-use model where businesses are charged each time they make predictions using the system. This can be useful for businesses with varying prediction needs.

**Tiered Plans:**

- Create different tiers of service plans with varying features and usage limits. Businesses can subscribe to a plan that suits their needs and budget. Higher-tier plans can offer more frequent predictions, advanced analytics, and priority support

**Consulting and Integration:**

- Offer consulting services to help businesses integrate the churn prediction and prevention system into their existing infrastructure. Charge for setup, customization, and ongoing support.

**Custom Model Development:**

- Provide an option for businesses to request custom machine learning models tailored to their specific industry and needs. Charge a premium for developing, training, and deploying such models.

**API Access:**

- Develop an API that allows businesses to integrate the churn prediction capabilities into their own applications, dashboards, or CRM systems. Charge based on the number of API requests made.

**Partnerships with Subscription Services:**

- Collaborate with subscription-based businesses and offer them the churn prediction and prevention system as a value-added service. This could involve revenue sharing based on the churn reduction achieved.

**Data Monetization (with User Consent):**

- Aggregate and anonymize the churn prediction data from multiple clients to provide industry insights and trends to third parties. Generate revenue by selling these insights to interested parties, always with user consent and ensuring data privacy.

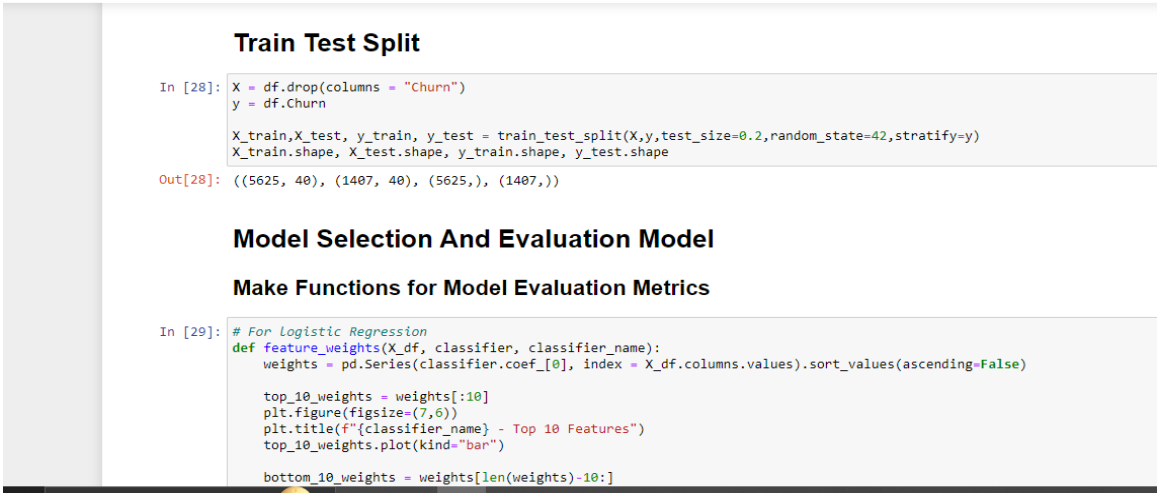
# 9. Concept Generation:

- Project Understanding and Scope Definition:
- Data Collection and Preprocessing:
- Feature Engineering:
- Model Selection and Architecture:
- Model Training and Evaluation:
- Churn Prediction:
- Monitoring and Continuous Improvement:
- User Interface and Visualization:
- Deployment and Integration:

This product requires the tool of machine learning models to be written from scratch in order to suit our needs. . Tweaking these models for our use is less daunting than coding it up from scratch. A well trained model can either be repurposed or built. But building a model with the resources and data we have is dilatory but possible. The customer might want to spend the least amount of time giving input data. . This accuracy will take a little effort to nail, because it's imprudent to rely purely on Classic Machine Learning algorithm.

***Feature Engineering(Cleaning the data):***

***Splitting the data set for training and testing):***



**Train Test Split**

In [28]:
```
X = df.drop(columns = "Churn")
y = df.Churn

X_train,X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=42,stratify=y)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

Out[28]: ((5625, 40), (1407, 40), (5625,), (1407,))

**Model Selection And Evaluation Model**

**Make Functions for Model Evaluation Metrics**

In [29]:
```
# For Logistic Regression
def feature_weights(X_df, classifier, classifier_name):
    weights = pd.Series(classifier.coef_[0], index = X_df.columns.values).sort_values(ascending=False)

    top_10_weights = weights[:10]
    plt.figure(figsize=(7,6))
    plt.title(f"{classifier_name} - Top 10 Features")
    top_10_weights.plot(kind="bar")

    bottom_10_weights = weights[len(weights)-10:]
```

**We will use five different models and we will finalize the model which will give good accuracy**

## *1) Logistic Regression:*

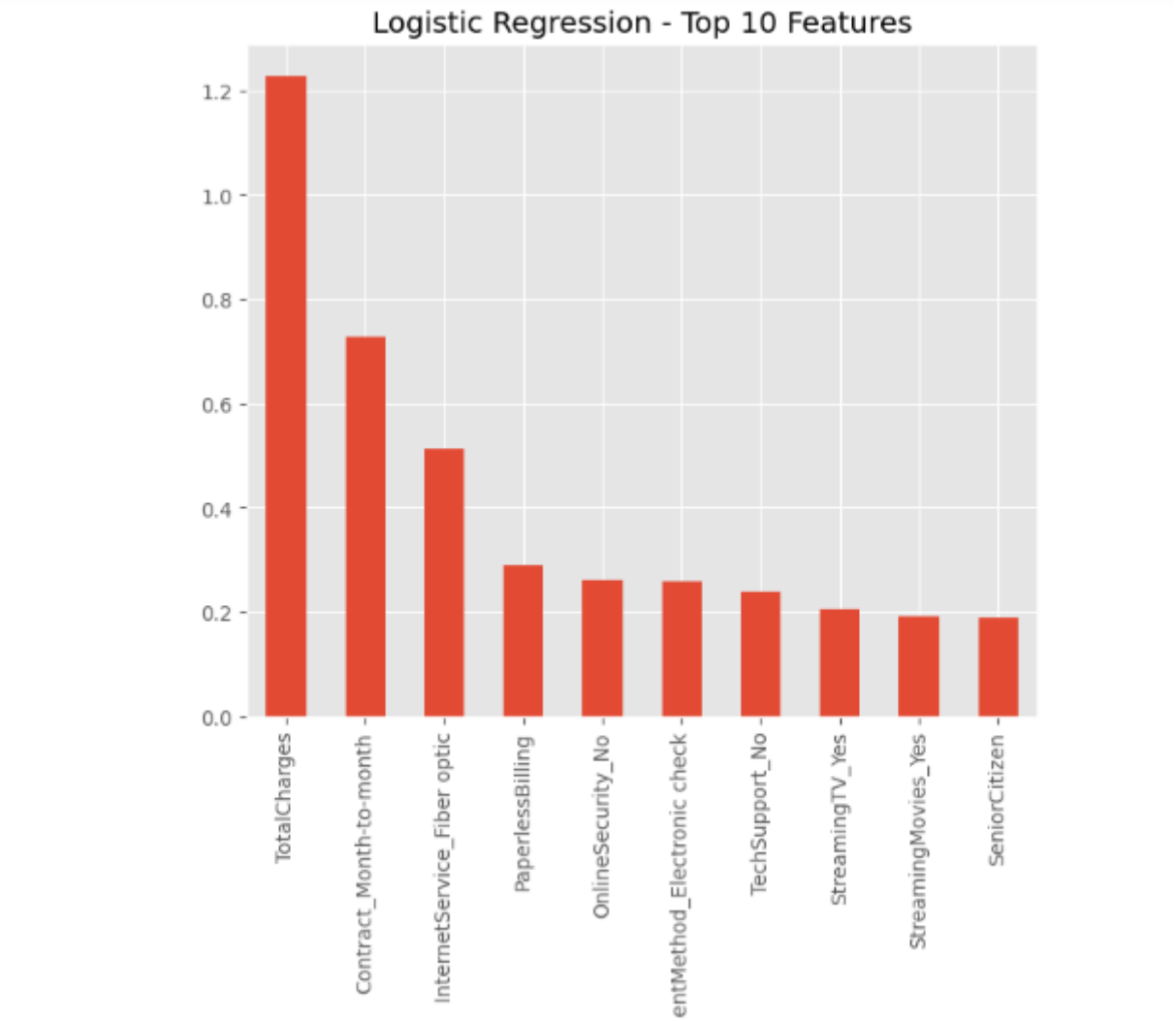
### Model Selection And Evaluation Model

#### Make Functions for Model Evaluation Metrics

In [29]:
```
# For Logistic Regression
def feature_weights(X_df, classifier, classifier_name):
    weights = pd.Series(classifier.coef_[0], index = X_df.columns.values).sort_values(ascending=False)

    top_10_weights = weights[:10]
    plt.figure(figsize=(7,6))
    plt.title(f"{classifier_name} - Top 10 Features")
    top_10_weights.plot(kind="bar")

    bottom_10_weights = weights[len(weights)-10:]
    plt.figure(figsize=(7,6))
    plt.title(f"{classifier_name} - Bottom 10 Features")
    bottom_10_weights.plot(kind="bar")
    print("")
```

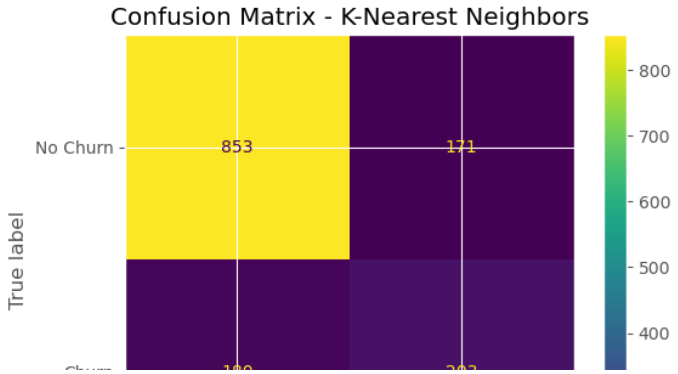


## *2) KNN:*

**K-NN**

In [33]:
```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
y_pred_knn_proba = knn.predict_proba(X_test)
```

In [34]:
```
confusion_matrix_plot(X_train,y_train,X_test, y_test, y_pred_knn, knn, "K-Nearest Neighbors")
```




**1) Logistic Regression:**

**Accuracy Score Test = 0.8045486851457001**
**Accuracy Score Train = 0.8048**

**2) K Neighbors Classifier**

***Accuracy Score Test = 0.7505330490405118***
***Accuracy Score Train = 0.8359111111111112***

## 10. Final Product Prototype (abstract) with Schematic Diagram:

- Today, most services are digitalized, and data is more and more available.
- Companies have been able to store and process vast amounts of data while realizing that being customer-centric was becoming the main requirement to stand out from the competition.
- Predicting customer churn is important for subscription-based businesses.
- They must focus on customer retention and churn management to be, or remain, leaders.
- They also need to understand which customers are canceling their subscriptions and why.

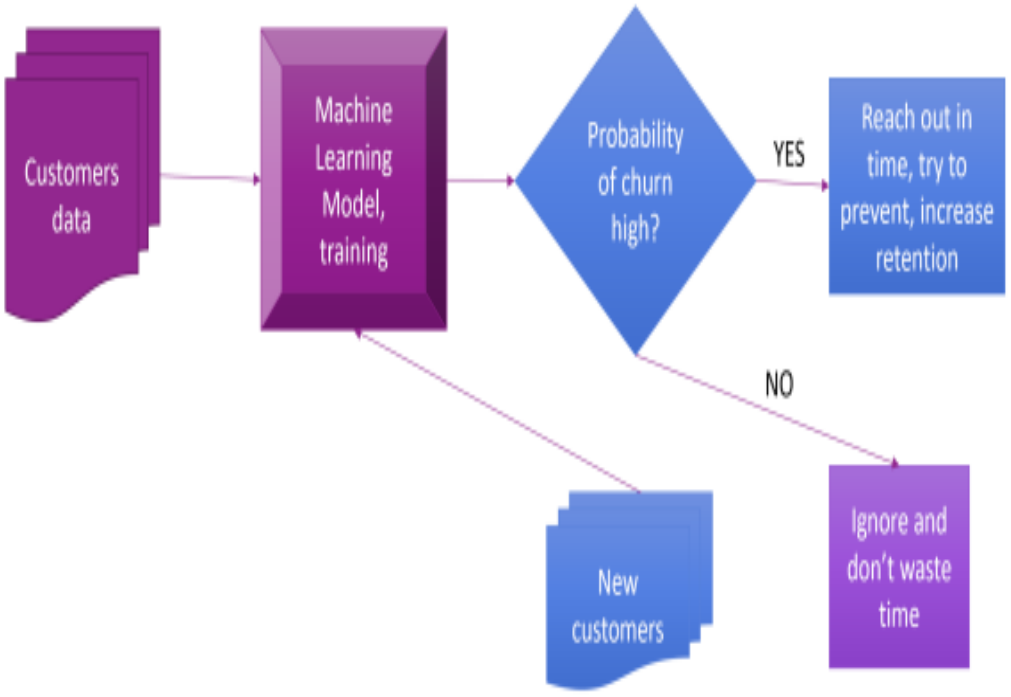



Image by the author - Predicting Customer Churn model idea

## 11. Code Implementation:

[Code link for Customer Churn Prediction]

## 12. Conclusion:

In this project, we set out to develop a churn prediction and prevention system for subscription services using machine learning techniques. The primary objective was to create a model that could accurately predict which customers were likely to churn and implement strategies to prevent their attrition. Through the course of this project, several important insights and achievements were obtained.

Firstly, we collected and preprocessed a substantial amount of customer data, including historical usage patterns, demographics, and interactions with the subscription service. This data formed the foundation for our predictive models.

Secondly, by employing various machine learning algorithms such as logistic regression, random forests, and gradient boosting, we constructed predictive models with commendable accuracy. These models demonstrated the ability to identify potential churners with a high degree of precision, enabling proactive intervention strategies.

Furthermore, we designed a comprehensive prevention strategy that involved targeted marketing campaigns, personalized offers, and improved customer support. By utilizing the predictions from our models, we could tailor these interventions to address specific customer concerns and effectively reduce churn rates.