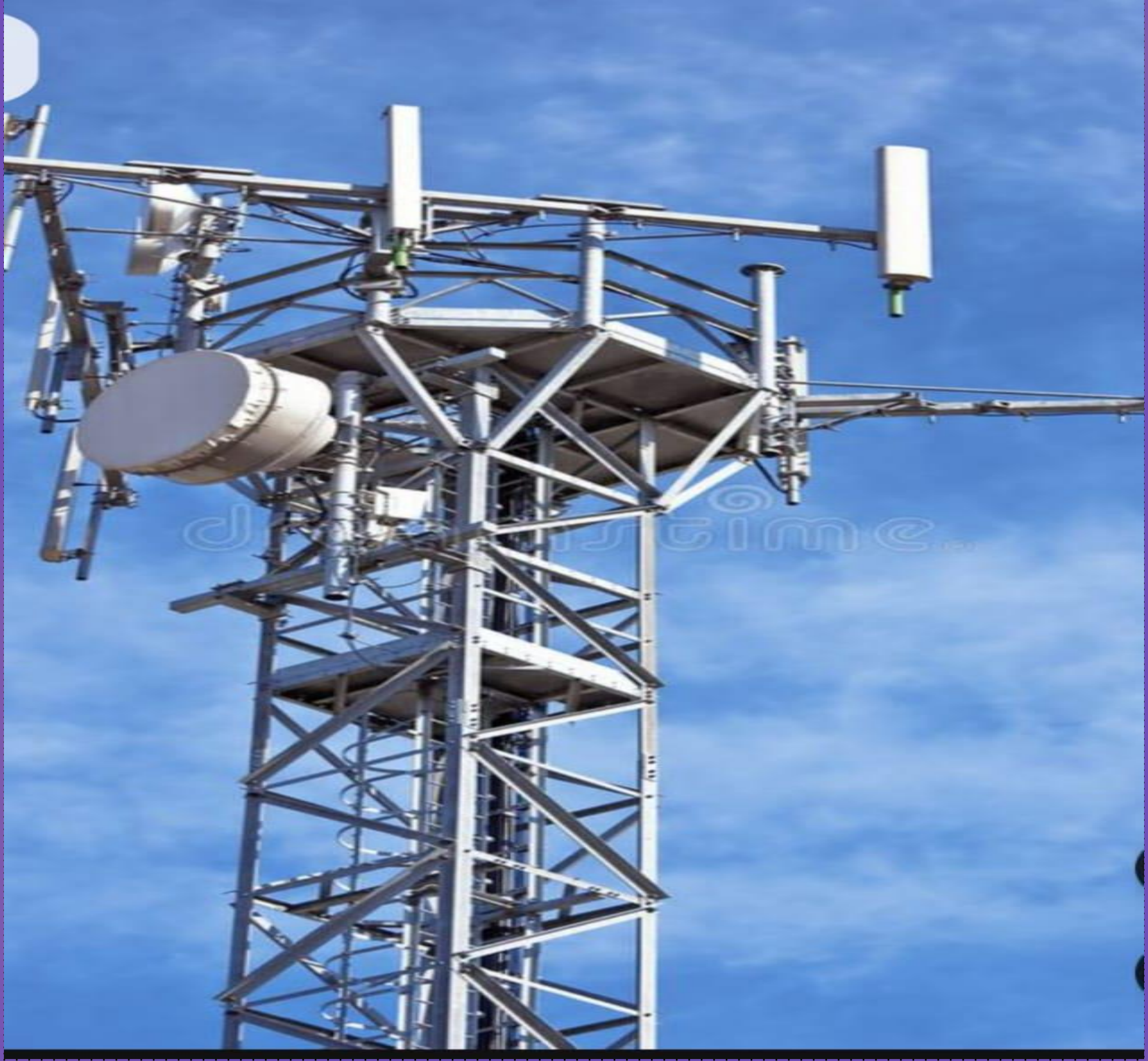


PROJECT OVER VIEW

THIS PROJECT AIM TO BUILD A MODEL THAT WILL PREDICT WHETHER A CUSTOMER IS LIKELY TO LEAVE SYRIATEL SOON

DATA UNDERSTANDING

- I conducted the following activities in my data understanding:

  1.Importing the relevant libraries

  2.Load the csv data set

  3.Checking the head and tail

  4.Checking the shape

  5.Checking the data types

  6.Checking the statistical summary

  7.Checking the columns

MODELING

- I was able to apply two classification models that is:

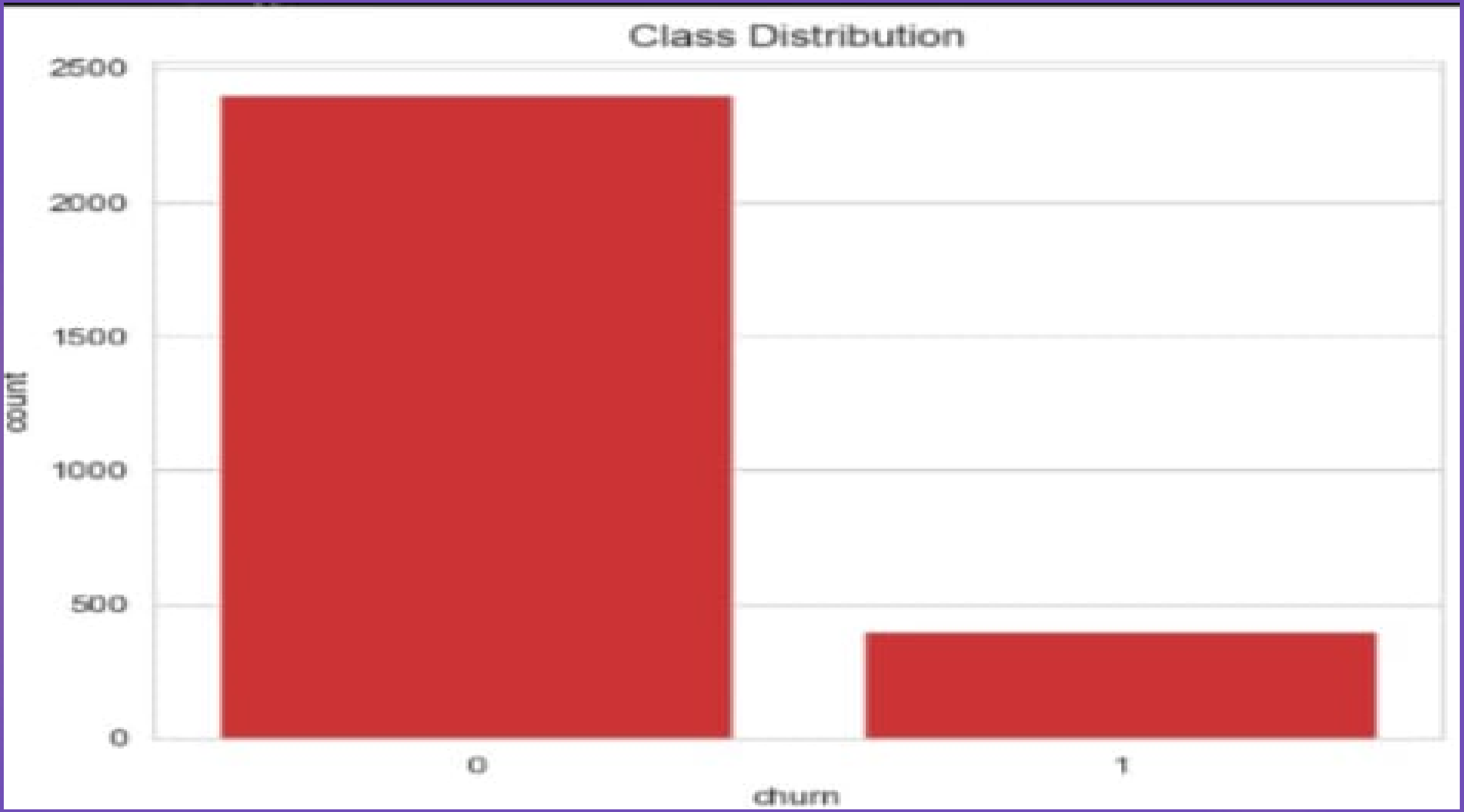
1. Logistic regression
2. Decision Trees

1. LOGISTIC REGRESSION

- I was able to do logistic regression when :

1. the class was not balanced
2. when the class was balanced

LOGISTIC REGRESSION WHEN THE CLASS IS NOT BALANCED



LOGISTiC REGRESSION WHEN THE CLASS IS NOT BALANCED(IMBALANCED)

- I conducted the following activities:

1. Preprocessing- I will Convert categorical variables (international plan, voice mail plan) into numerical format.

   * I will Convert churn (target variable) into a binary numerical format (False → 0, True → 1).

   * I will also drop state because it has too many categorical variables

LOGISTiC REGRESSION WHEN THE CLASS IS NOT BALANCED(IMBALANCED)

2. Splitting the data into target and Predictor. The target variable is churn while the others are predictor variables

3. Splitting the data into train,test and split.I used a test size of o.1 and a random state of 42 because they give the highest accuracy score

4.Scalling. I performed scalling on the features and not the target variables

5.Prediction.I predicted on the y based on the x_test

6.Accuracy. I checked on the accuracy and got an accuracy of 89 %

LOGISTiC REGRESSION WHEN THE CLASS IS BALANCED(SMOTE TECHNIQUE)

- I used the smote technique to address the issue of class imbalance
- I went ahead to conduct all activities that include:

1. Preprocessing- I will Convert categorical variables (international plan, voice mail plan) into numerical format.

   * I will Convert churn (target variable) into a binary numerical format (False → 0, True → 1).

   * I will also drop state because it has too many categorical variables

LOGISTiC REGRESSION WHEN THE CLASS IS BALANCED(SMOTE TECHNIQUE)

2. Splitting the data into target and Predictor. The target variable is churn while the others are predictor variables

3. Splitting the data into train,test and split.I used a test size of 0.1 and a random state of 42 because they give the highest accuracy score

4.Scalling. I performed scalling on the features and not the target variables

5.Prediction.I predicted on the y based on the x_test

6.Accuracy. I checked on the accuracy and got an accuracy of 49 %

2. DECISION TREES USING HYPERPARAMETER

- Here I used max_depth as a tuned hyper parameter
- I was able to conduct the processes that include the following:

1. Preprocessing- I will Convert categorical variables (international plan, voice mail plan) into numerical format.

   * I will Convert churn (target variable) into a binary numerical format (False → 0, True → 1).

   * I will also drop state because it has too many categorical variables

DECISION TREES USING HYPERPARAMETER

2. Splitting the data into target and Predictor. The target variable is churn while the others are predictor variables

3. Spliting the data into train, test and split. I used a test size of 0.1 and a random state of 42 because they give the highest accuracy score

4.Scaling. I performed scaling on the features and not the target variables

5.Prediction.I predicted on the y based on the xtest

6.Accuracy. I checked on the accuracy and got an accuracy of 86 %

7.Visualization.I used a max_depth of 6

EVALUATION

- Here I was able to determine which is the best model using the accuracy score and confusion matrix

1. Accuracy score.The best performing model was Decision trees because it had accuracy score of 86% compared to 49% by logistic regression(when the class was balanced)

2. Confusion matrix. Based on the confusion matrix decision tree is the best as we compare the number of true positives values

RECCOMENDATIONS

- Based on the models above here is a recommendation:

- Both models (decision tree and logistic) have the tendency to predict the most frequent class. By using methods like oversampling the less frequent class e.g using SMOTE, undersampling the frequent class, or adjusting class weights, the models may be capable of learning patterns of the less frequent class better.

NEXT STEP

- Here is the next step of action:
- Consider using other method such as Random Forests as they often handle imbalance more robustly by combining multiple models.

THANK YOU