TECHNOCOLBS DATA SCIENCE INTERNSHIP

JULY 2021

MAJOR PROJECT REPORT

TITLE: **Predicting Flights Delay Using Local Weather Data**

ABSTRACT:

Flight delay is inevitable and it plays an important role in both profits and loss of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and incomes of airline agencies. The primary goal of this project is to predict airline delays caused by various factors. Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions. Hence, these factors indicate how necessary and relevant it has become to predict the delays no matter the wide-range of airline meshes. To carry out the predictive analysis, which encompasses a range of statistical techniques from supervised machine learning and, data mining, that studies current and historical data to make predictions or just analyze about the future delays, with help of Regression Analysis using regularization technique in Python This prediction will be helpful for giving a detailed analysis of the performance of individual airlines, airports, and then making a well-assessed decision. Moreover, apart from the assessment related to the passengers, delay prediction analysis will also help in important decision-making procedures necessary for every pivotal player in the air transportation system. There have been many researches on modeling and predicting flight delays, where most of them have been trying to predict the delay through extracting important characteristics and most related features. However, most of the proposed methods are not accurate enough because of massive volume data, dependencies and extreme number of parameters. Our approach was to clean the data also find the relevant insights from the data and then use supervised learning algorithm such as logistic regression, Decision Tree Classifier and Linear Discriminant Analysis (LDA) to Predicting the Flights Delay with prediction accuracy of 83%.

INTRODUCTION:

Flight Delay prediction is highly based on the weather forecast of the place and the changes it has during the day. Moreover, the different seasons lead to more air traffic. We started with treating the NULL VALUES in the data and since it is a massive dataset, we removed the features which had more than 45% null values. Then we analyzed the data using EDA (Exploratory Data Analysis) which included using different visualization like Histogram plot, Heatmap etc. Here we used Supervised learning.

Supervised Learning:

A Supervised Machine Learning It is a machine learning task where the dataset inputs and outputs are clearly recognized and already given, then several type of algorithms are trained using labeled examples. A supervised learning algorithm contains an entire dataset, which is further divided into training and test data; the algorithm examines the training dataset and produces an inferred function, which is then used for mapping new examples. Supervised Learning algorithm here will model relationships and dependencies between the aimed prediction output and the input features, such that I'll be predicting the output values for new data based on the relationships which are learned from the previous data set. Supervised Learning problems can be further categorized into following problems.

Classification:

It is a type problem in which the output variable is an entire category itself, such as "Win" or "Lose", the entire input data is classified into the category variables; it is generally used largely for recommendation problems Regression – It is a type of problem is which the output variable is a real value, such as few raw data values related to something. This is the problem type massively used for prediction analysis, and hence will be used in this project. B. Regression Analysis Methods The main focus of regression analysis is to model and determine the expected value of a dependent variable y in terms of the value of one or more independent variables (x). •

Linear Regression Linear Regression is used to model and establish a relationship between dependent and independent set of variables by fitting the best line possible. Hence formed as the result of prediction carried out is known as our regression line and is represented by a linear equation (1) y = b0+b1x1 .
In case of logistic regression, which is very much compared to linear regression, the outcome (dependent variable) has only limited number of discrete possible values. Whereas, linear regression analysis is the first best suited method because it results in any one among the range of an infinite number of possible values.

Flight Delays has become a common and complex phenomenon, it occurs due to the problems at the origin airport, at the destination-airport, any ground reasons or a combination of these entire factors can also give rise to delays. Delays are also being regarded as caused due to specific airlines. Even if it is complex, it is still measurable with decent accuracy. And with respect to the schedule and on-time performance of airlines, their generally exists some pattern of flight delay. The results obtained from this project, Airline Delay Predictions using Supervised Machine Learning, it can help to better understand the phenomenon and up to a very large extent.

DATA:

The dataset has been taken from a reliable online available government agency website that provides the air traffic delay statistics in the United States. The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. BTS compiles daily data for the benefit of the customers or for any data analysts. The dataset is of 2017 flight delays and cancellations. Fig.2: All the airlines in the dataset associated with particular IATA carrier codes. B. Data Exploration Data cleaning is the critical initial step in evaluating the dataset for final analysis. With the enormous amount of data available, databases are prone to have noisy, missing and inconsistent data. The data in this project is obtained from TABLE I DESCRIPTION OF THE ATTRIBUTES INVOLVED IN THE DATASET Attributes Descriptions of Attributes YEAR, MONTH, DAY, DAY_OF_WEEK dates of the flight AIRLINES ORIGIN_AIRPORT and DESTINATION_AIRPORT SCHEDULED_DEPARTUR E and SCHEDULED_ARRIVAL DEPARTURE_TIME and ARRIVAL_TIME DEPARTURE_DELAY and ARRIVAL_DELAY DISTANCE, which has varying kinds of 31 variables involved, and may not be compatible with the format in which we require the data to use in Python. Data Cleaning helps in removing noisy data, and removing inconsistencies. Data cleaning is performed as follows:
Dates and Times: The date format has been given in four variables format; it will be toned down to one particular format available in Python for ease of use.

```
flight_df['DATE'] = ''
flight_df['DATE'] = pd.to_datetime(flight_df[['YEAR', 'MONTH', 'DAY']])
```

Filling Factor: In the data cleaning process, a missing value can be ignored, manually entered, given a constant value, or a mean value. In this case, it will be organizing and arranging the entire data frame to keep the relevant attributes and eliminate the ones which has missing values. This is done to increase the readability and feasibility of use.

```
#FILLNA USING COLUMNS MEAN
#flight_df.fillna(flight_df.mean(),inplace=True)
```

```
flight_df['DEPARTURE_TIME'].fillna(flight_df['SCHEDULED_DEPARTURE'],inplace=True)
```

```
flight_df['DEPARTURE_DELAY'].fillna(flight_df['DEPARTURE_DELAY'].mean(),inplace=True)
```

```
flight_df['TAXI_OUT'].fillna(flight_df['TAXI_OUT'].mean(),inplace=True)
```

```
flight_df['WHEELS_OFF'].fillna(flight_df['WHEELS_OFF'].mean(),inplace=True)
```

```
flight_df['SCHEDULED_TIME'].fillna(flight_df['SCHEDULED_TIME'],inplace=True)
```

```
flight_df['ELAPSED_TIME'].fillna(flight_df['ELAPSED_TIME'].mean(),inplace=True)
```

```
flight_df['AIR_TIME'].fillna(flight_df['AIR_TIME'].mean(),inplace=True)
```

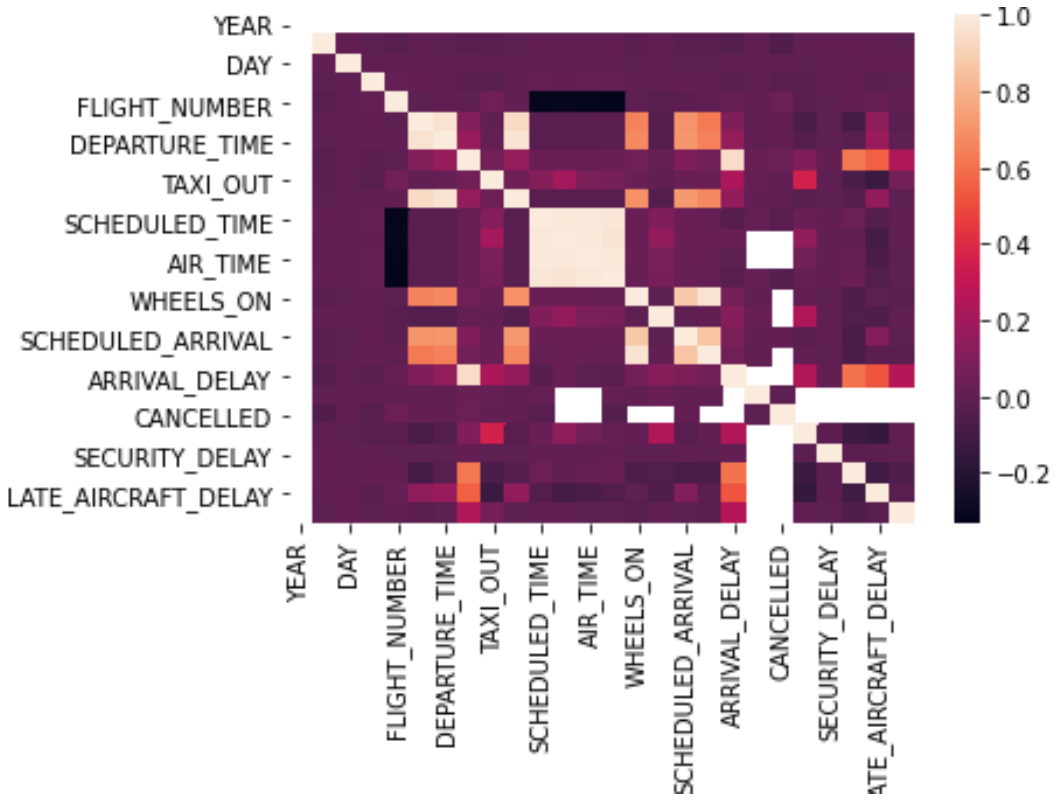
```
flight_df['WHEELS_ON'].fillna(flight_df['WHEELS_ON'].mean(),inplace=True)
```

```
flight_df['TAXI_IN'].fillna(flight_df['TAXI_IN'].mean(),inplace=True)
```

```
flight_df['ARRIVAL_TIME'].fillna(flight_df['SCHEDULED_ARRIVAL'],inplace=True)
```

```
flight_df['ARRIVAL_DELAY'].fillna(flight_df['ARRIVAL_DELAY'].mean(),inplace=True)
```

Then we used Heatmap to see the multicollinearity in the data.



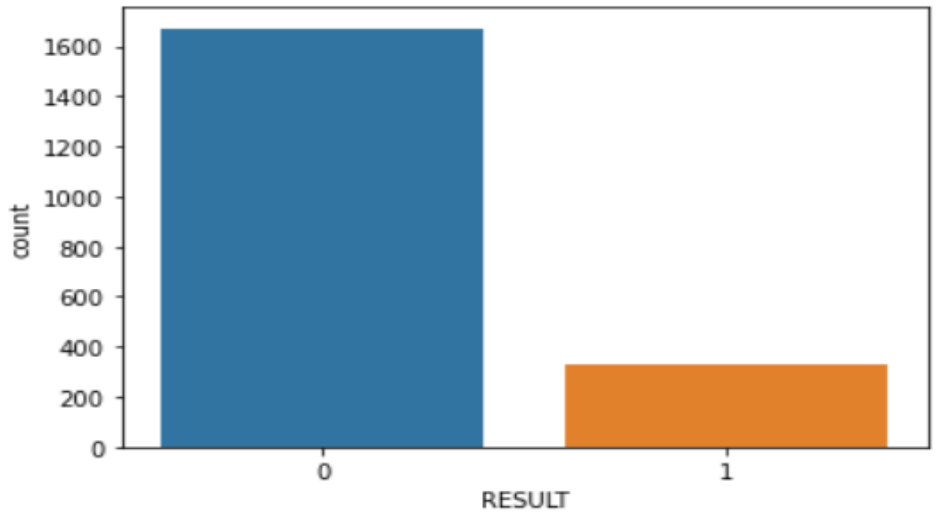


Fig4: *The Count Chart shows, Not Delay Count: 83.0, Delay Count: 16.0.*

MODEL BUILDING:

Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look. The learning is derived from data. The right machine learning approach and methodologies stem from data-centric needs and result in projects that focus on working through the stages of data discovery, cleansing, training, model building and iteration.

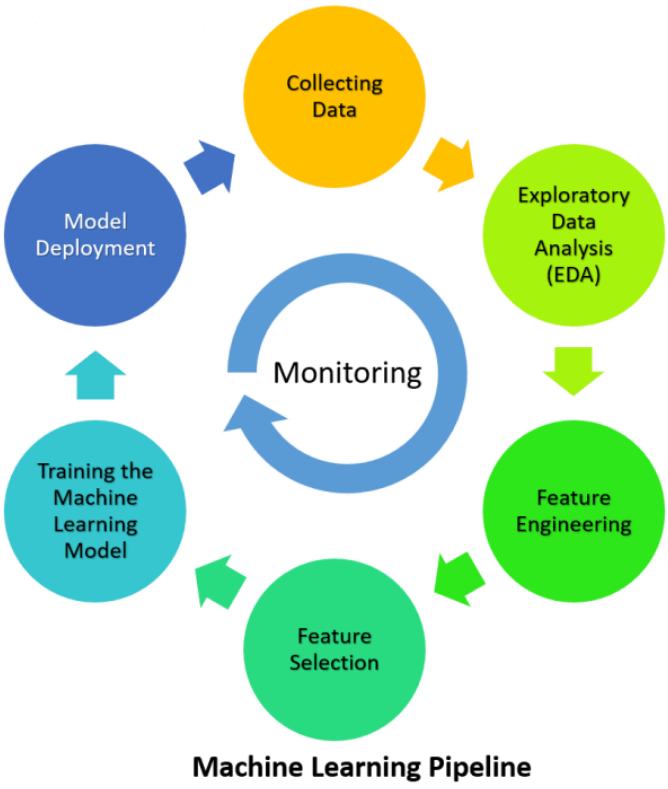


Fig6: *The Model Building chart show us the steps that involve in the model development*

Logistic Regression

```
from sklearn.linear_model import LogisticRegression
Lr=LogisticRegression()
Lr.fit(X_train_sc,y_train)
```

```
78]: LogisticRegression()
```

Train Model

```
y_pred_trainlr1 =Lr.predict(X_train_sc)
```

```
y_pred_trainlr1
```

```
0]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

```
Lr.score(X_train_sc,y_train)
```

```
1]: 0.8378571428571429
```

Fig7: *With the Logistics Regression to predict the flights delay based on the Local weather data*

Now when we tested the model, we got following result.

Test Model

```
In [85]: y_pred_testlr1 =Lr.predict(X_test_sc)
```

```
In [86]: y_pred_testlr1
```

```
In [87]: Lr.score(X_test_sc,y_test)
Out[87]: 0.8233333333333334
```

```
In [88]: confusion_matrix(y_test, y_pred_testlr1)
Out[88]: array([[494,   0],
                [106,   0]], dtype=int64)
```

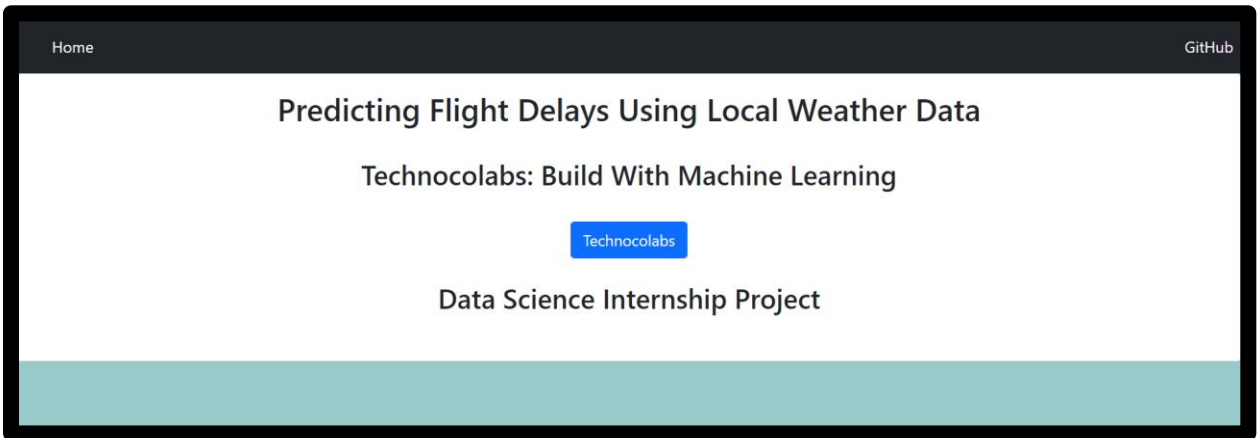
```
In [89]: print(classification_report(y_test, y_pred_testlr1))
```

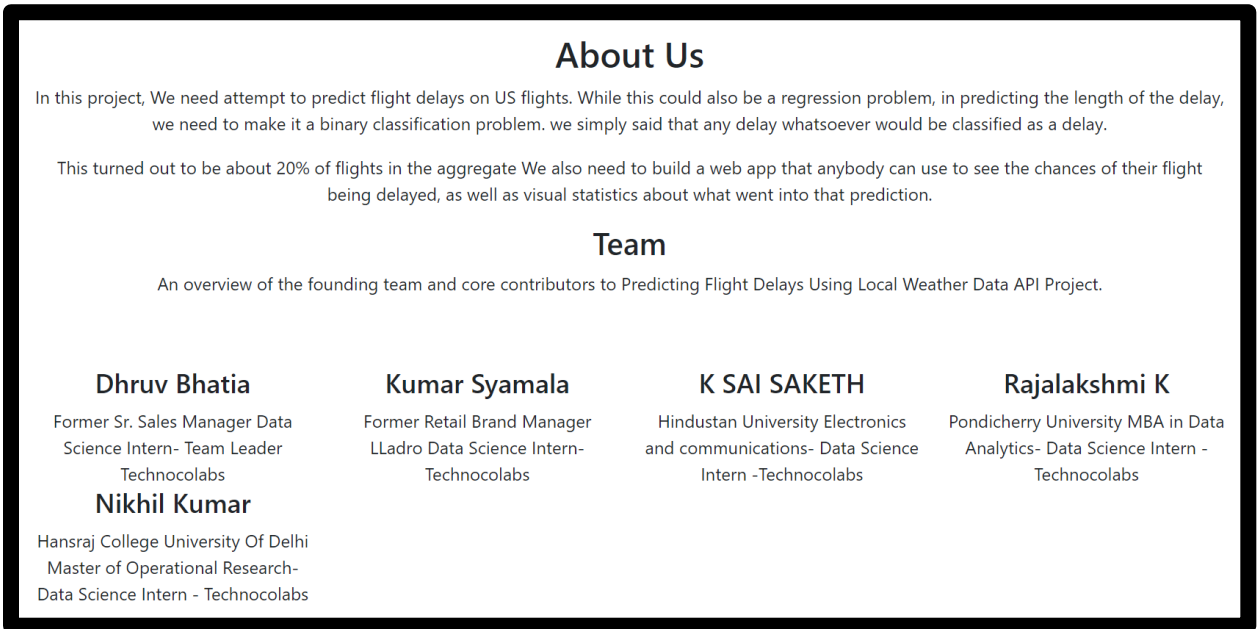
	precision	recall	f1-score	support
0	0.82	1.00	0.90	494
1	0.00	0.00	0.00	106
accuracy			0.82	600
macro avg	0.41	0.50	0.45	600
weighted avg	0.68	0.82	0.74	600

DEPLOYMENT

DEMO:

API Link: https://predictingflightsdelay-api.herokuapp.com/





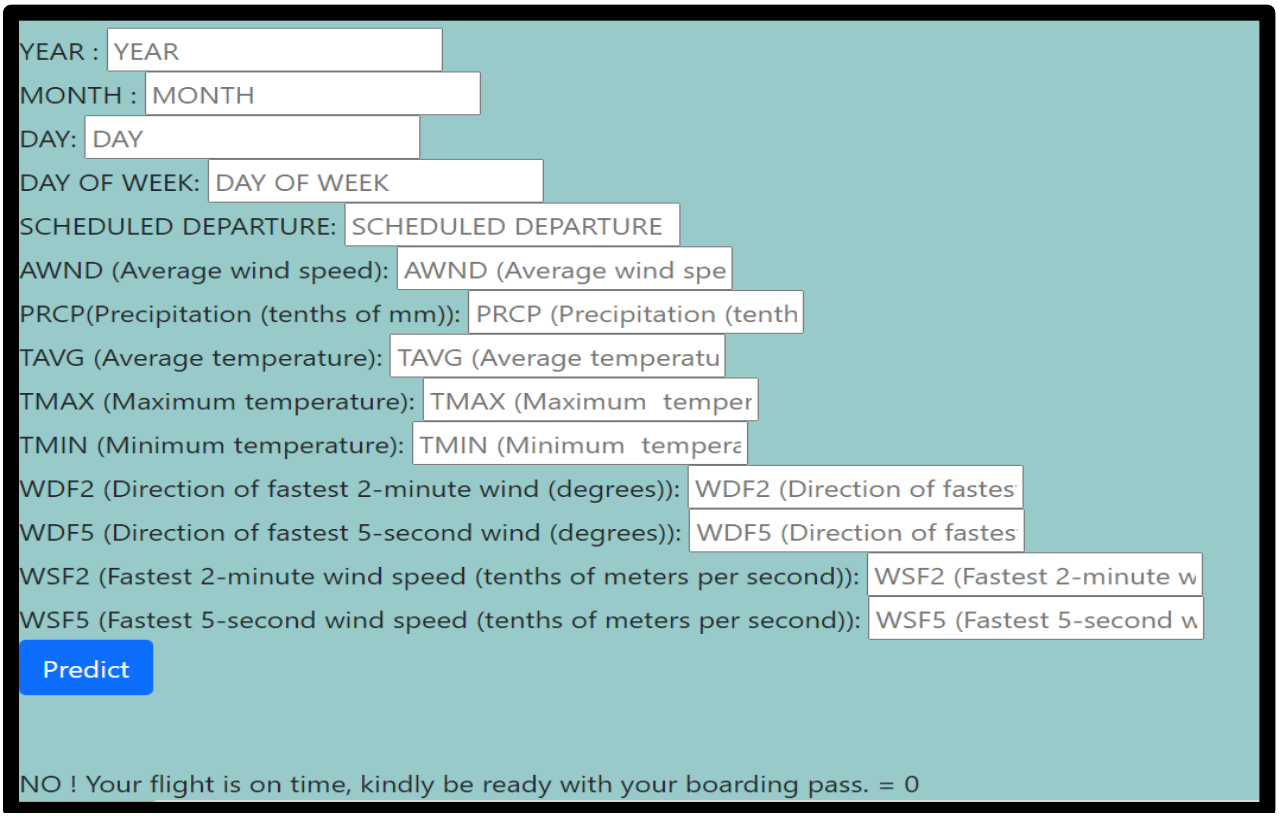


Fig8: *The API Overall Outlook. Frontend developed with the usage of HTML/CSS/JS/Bootstrap. The Result will show the predicted price in form of [0]: Not Delay or [1]: Delay.*

OVERVIEW:

This is a Flask web app which predicts the bitcoin price based on the twitter sentiments.

INSTALLATION:

The Code is written in Python 3.9.0. To install the required packages and libraries, run this command in the project directory after cloning the repository:

```
pip install -r requirements.txt
```

DEPLOYMENT ON HEROKU:

Login or signup in order to create virtual app. This can be done either connect GitHub profile or download ctl to manually deploy this project.

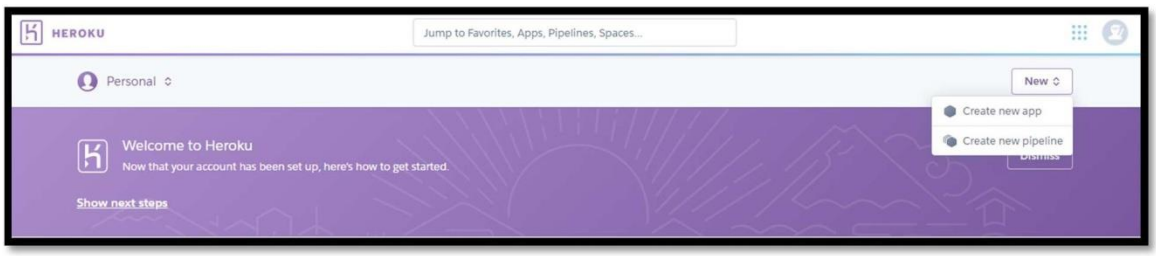


Fig9: *Heroku Login Page Overview: Next step would be to follow the instructions given on Heroku Documentation (https://devcenter.heroku.com/articles/getting-started-with-python) to deploy a web app.*

DIRECTORY TREE:

```
/Model-prediction
├── templates
│   └── index.html
├──app.py
├──Procfile
├── requirements.txt
├──gitignore
├──model.pkl
```

TECHNOLOGIES USED:

TEAM MENTOR:

DHRUV BHATIA
PROJECT LEAD, TECHNOCOLABS

TEAM MEMBERS:

1. S. KUMAR RAMA KOTI REDDY
2. NIKHIL KUMAR
3. K SAI SAKET
4. RAJALAKSHMI K

****