Customer Churn Analysis

By Prakhar Chitransh

Dataset—https://www.kaggle.com/datasets/blastchar/telco-customer-churn

For IVYKIDS

ABSTRACT

In the ever-changing corporate environment of today, it is critical to comprehend and reduce client attrition in order to maintain profitability and steady growth. Using advanced machine learning techniques, this project explores the field of customer churn analysis in an effort to identify trends and forecast possible client attrition. As a conscientious student , My goal was to use cutting-edge methods to streamline decision-making processes and deliver actionable insights as mentioned by IVYKIDS.

Traditional decision tree models are first examined in the study, which yields insightful information on feature importance and decision paths. Recognizing the drawbacks of decision trees, however, the research goes a step further and uses Random Forest, a potent ensemble learning technique. Random Forest was found to regularly outperform decision trees through extensive experimentation and model evaluation, exhibiting higher accuracy rates and improved generalization.

Principal Component Analysis (PCA) was added to the feature engineering pipeline in order to enhance the analytical framework. PCA reduced dimensionality while maintaining the underlying patterns in the data, making it easier to find and extract important latent variables. PCA was used into the Random Forest model to improve interpretability and create a more reliable and effective churn prediction model.

This report encapsulates the journey of transitioning from conventional decision trees to the formidable Random Forest algorithm, emphasizing the pivotal role played by PCA in refining the predictive capabilities of the model. The findings of this study not only demonstrate the efficacy of advanced machine learning methodologies in customer churn analysis but also underscore the potential for improved decision support systems within the company's data science initiatives.

Combining Random Forest with PCA is a powerful way to improve the predictive models' accuracy and resilience, which will enable the business to better handle customer attrition and maintain its place in the market.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

PCA   Principal Component Analysis

DT   Decision Tree

RF   Random Forest

CHAPTER 1

INTRODUCTION

1.1 Overview

I have gone through the learning method of customer churn Analysis and how the customer behaviour changes with aspects to different situations in the dataset and filtered out the best possible outcome for each of it for this IVYKID project and generated a lot of insight through the Datasset that was provided to me.

1.2 Motivation

Understanding customer churn is not merely a statistical endeavor; it is a strategic imperative that can directly impact the bottom line. By delving into the patterns and indicators of customer attrition, IvyKids can gain a nuanced understanding of the factors influencing churn and, in turn, formulate targeted interventions to mitigate it. In the realm of education and child care, where personalized relationships and trust are paramount, retaining customers is not only a matter of financial prudence but also a testament to the quality and efficacy of the services provided.

1.3 Objective

The primary objective of the customer churn analysis project assigned by IvyKids is to leverage advanced data science techniques to gain profound insights into customer behavior and factors influencing churn within the education and child care domain. As a shortlisted candidate for the data scientist role, the overarching goal is to contribute meaningfully to IvyKids' strategic objectives by achieving the following specific objectives:

1.4 Scope

The project aims to analyze customer churn within the telecom industry using a comprehensive dataset that includes information about existing subscribers and potential subscribers.The focus is on understanding the factors contributing to churn and developing predictive models to identify and mitigate potential attrition.

CHAPTER 2
Explanation of the
Project

Data Collection and Preprocessing:

- Utilizing the provided dataset from Kaggle (link: Telco Customer Churn Dataset).

- Performing data preprocessing to address missing values, encode categorical variables, and ensure the dataset is ready for analysis.

Exploratory Data Analysis (EDA):

- Conducting a thorough exploration of the dataset to comprehend customer behavior and identify factors influencing churn.

- Utilizing statistical measures and visualizations (e.g., histograms, heatmaps, and correlation matrices) to communicate key findings effectively.

Feature Engineering:

- Developed relevant features that enhance the predictive capabilities of the model.

- Considered creating new variables or transforming existing ones based on insights gained from the EDA.

Building the Churn Prediction Model:

- Chose suitable machine learning algorithms for churn prediction, such as random forests, Decision Tree,PCA(Principal Component Analysis).

- Implemented and trained the selected models using the preprocessed dataset.

- Fine-tuned hyperparameters to optimize model performance.

Model Evaluation:

- Assessed the performance of the churn prediction models using appropriate evaluation metrics, including accuracy, precision, recall, and F1-score.

- Conducted a comparative analysis to identify the most effective model for deployment.

Data display of all the useful attributes-

Displaying Target variable per Category-

Percentage of Missing Values-
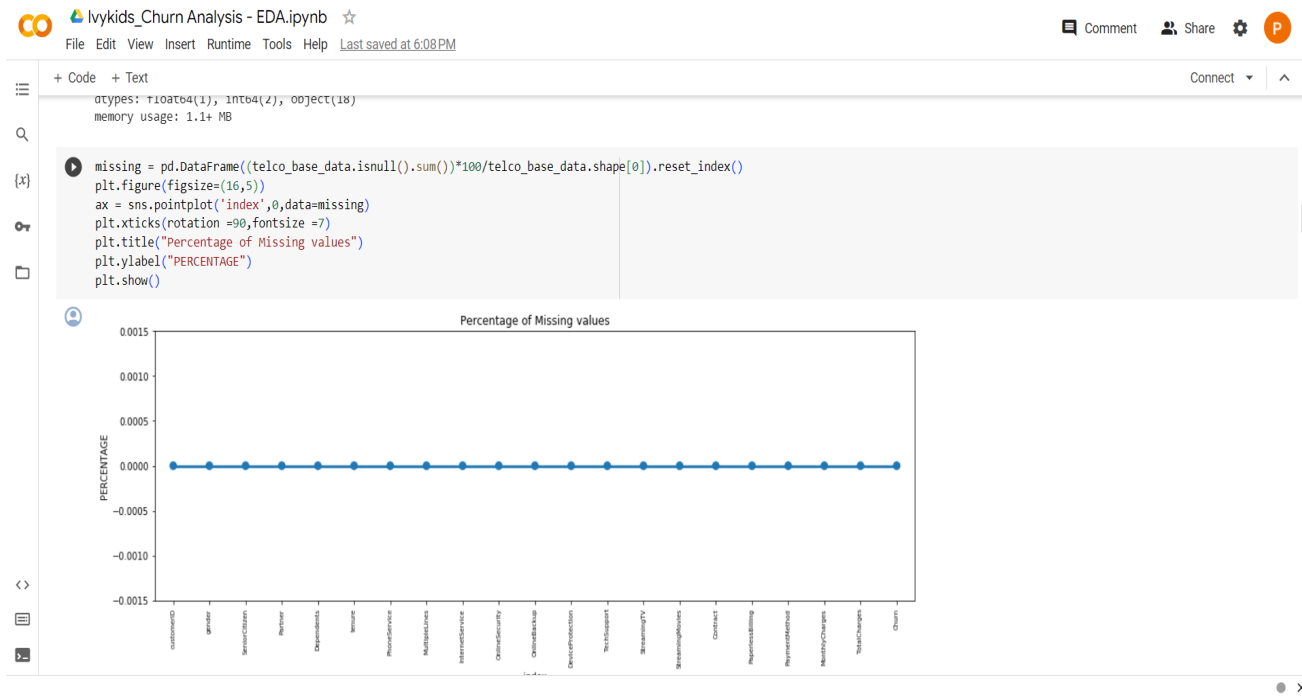
Churn Analysis of Senior Citizen-
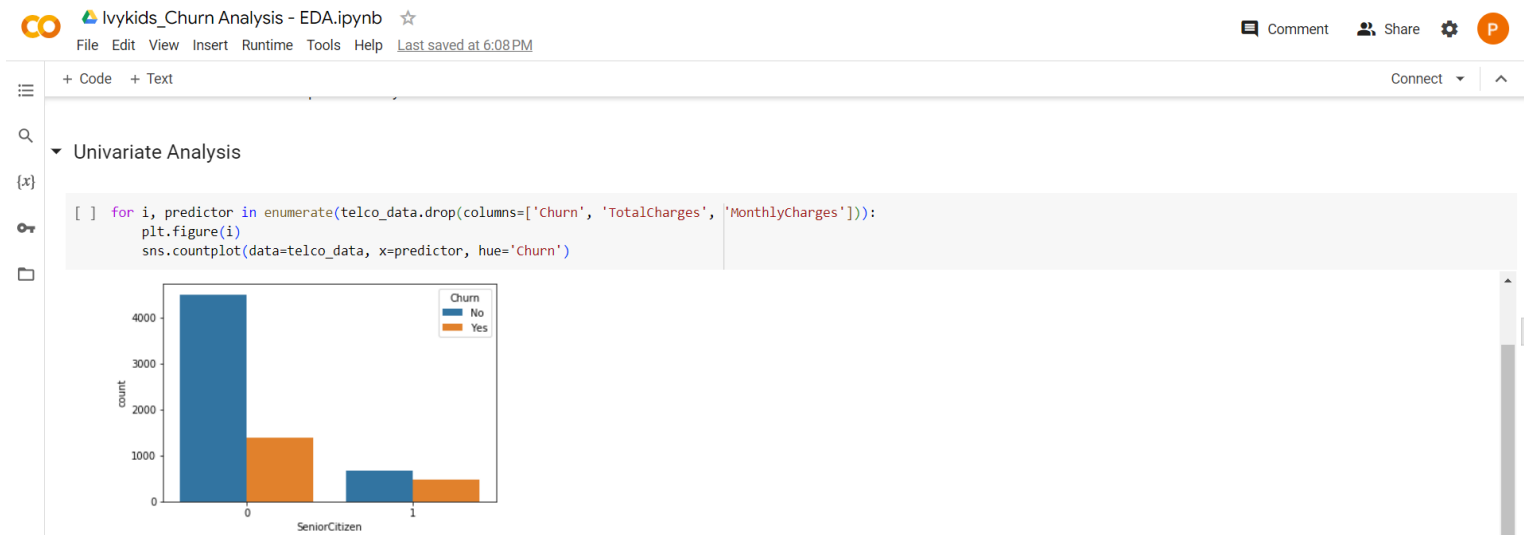
Churn Analysis of different useful attributes-

▾ Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```

▾ Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```

▾ Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```

Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```
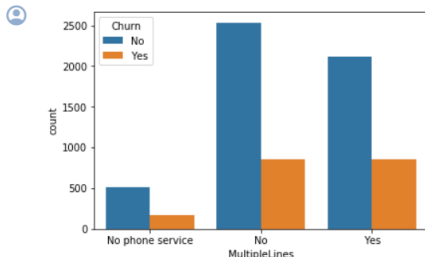
Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```

Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```
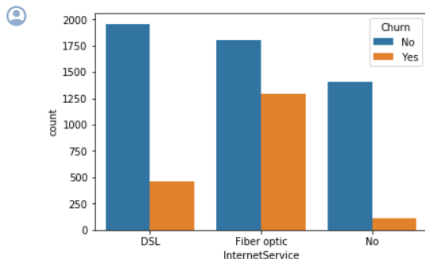
Univariate Analysis

```python
for i, predictor in enumerate(telco_data.drop(columns=['Churn', 'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=telco_data, x=predictor, hue='Churn')
```
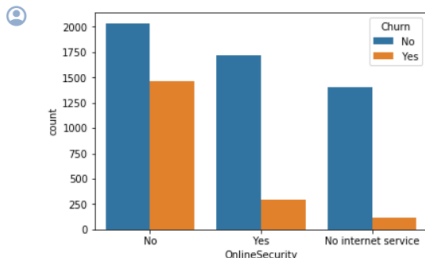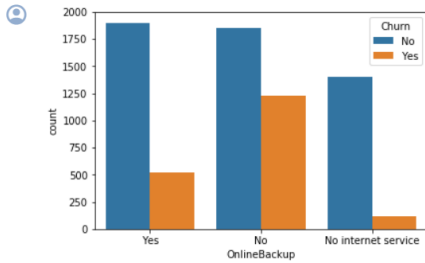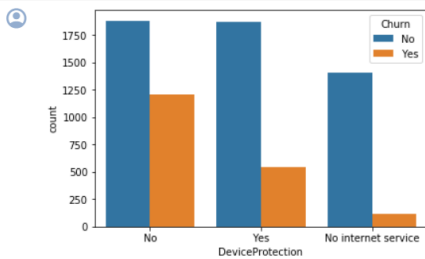
CHAPTER 3

PROPOSED METHODOLOGY

3.1 Dataset and Models

A. Dataset:

Numerous parameters, including customer demographics, contract information, usage trends, and customer satisfaction ratings, are included in the Telco Customer Churn dataset. "Churn," the goal variable, indicates whether or not a customer has left.

B. Models:

Several machine learning algorithms, including Logistic Regression, Decision Trees,PCA(Principal Component Analysis), and Random Forest,I have considered. Random Forest demonstrated the highest accuracy during model evaluation.

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. This ensemble approach helps to mitigate overfitting and improve the model's overall generalization performance.

Random Forest has exhibited higher accuracy compared to individual decision trees. By aggregating the predictions of multiple trees, it reduces the impact of outliers and noisy data, leading to a more robust and accurate model.

Decision trees, when grown deep, are prone to overfitting the training data, capturing noise rather than true patterns. Random Forest mitigates this by constructing multiple trees with random subsets of features, leading to a reduction in overfitting.

Random Forest provides a feature importance ranking, which is valuable in understanding the relative contribution of each feature to the model's predictions. This aids in interpreting the factors influencing customer churn and informs targeted business strategies.

Customer churn is often influenced by complex and non-linear relationships between various factors. Random Forest is well-suited for capturing these non-linearities, making it effective in scenarios where simple linear models or decision trees might fall short.

Decision cycle of subscriber for telecommunication subscriber-
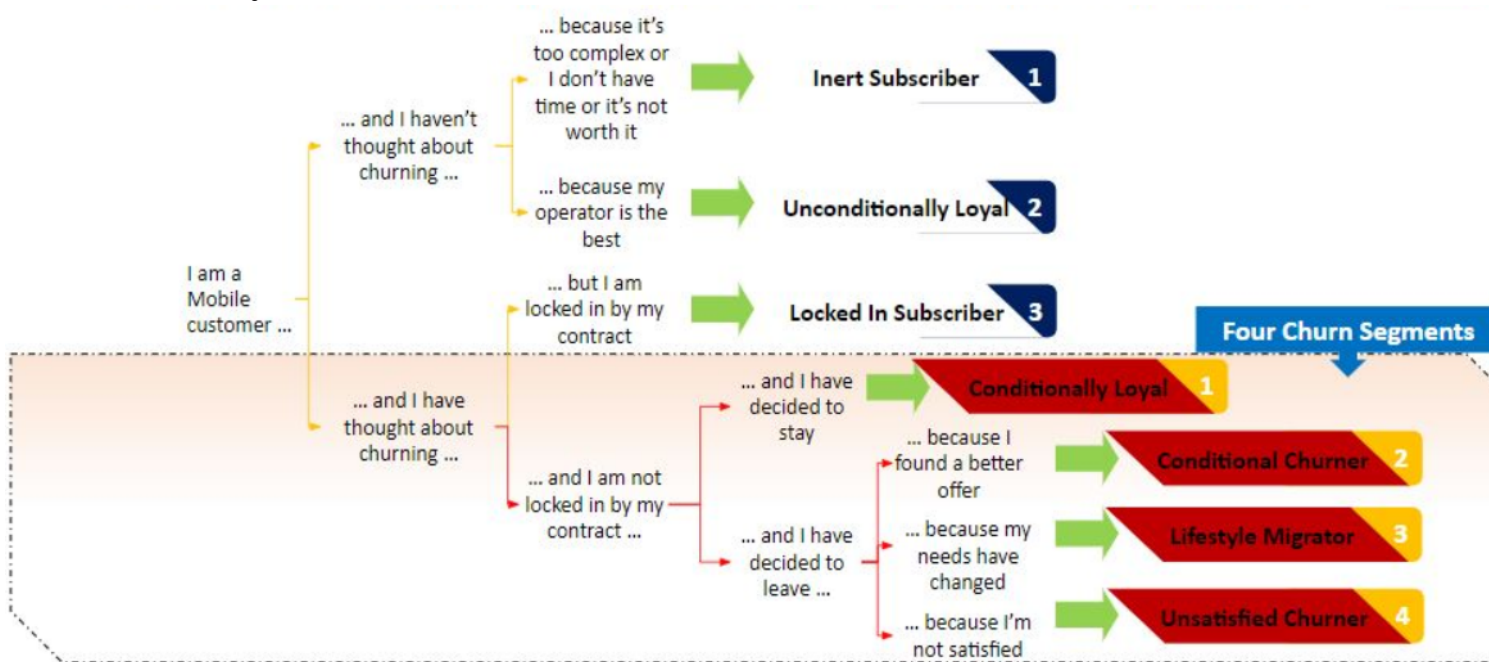
Heatmap of different attributes-
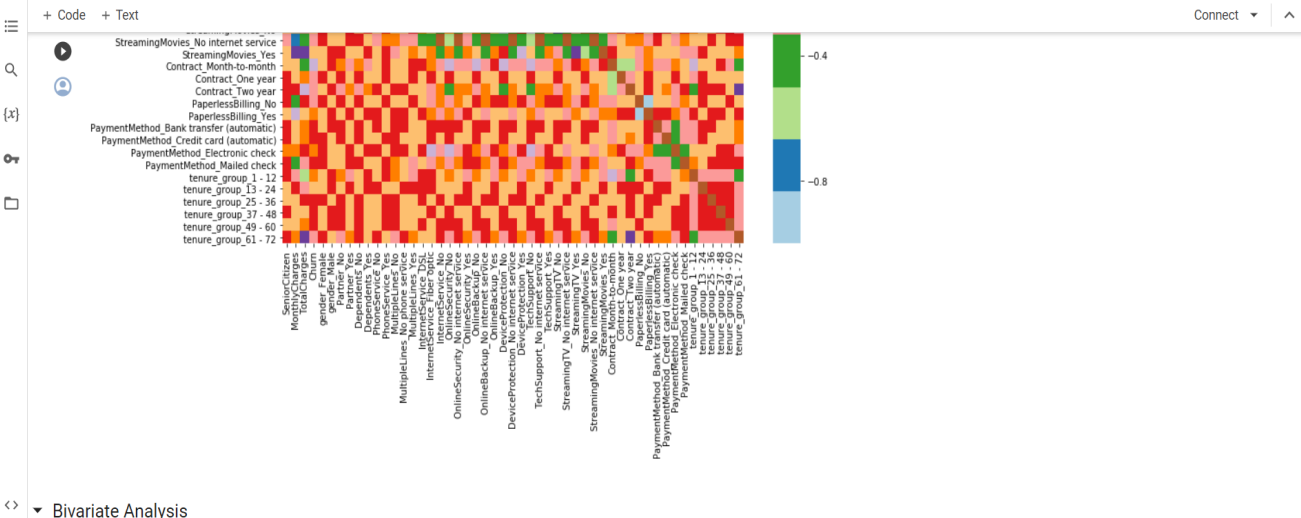
Bivariate Analysis

METHODOLOGY/WORKFLOW

Model Selection: Random Forest

Random Forest was chosen as the machine learning algorithm for its ensemble learning approach and proven success in predictive tasks. Ensemble methods, which combine the predictions of multiple models, often lead to better generalization performance compared to individual models.

Model Building:

The Random Forest model was implemented using Python's scikit-learn library. The features (X) and target variable (y) were defined, and the model was trained on the training dataset. The ensemble of decision trees within Random Forest allows for robust predictions while mitigating overfitting.

Hyperparameter Tuning:

Fine-tuning the model's hyperparameters was a critical step to optimize its performance. Techniques like grid search were employed to find the best combination of hyperparameters, ensuring that the model achieved its highest potential accuracy.
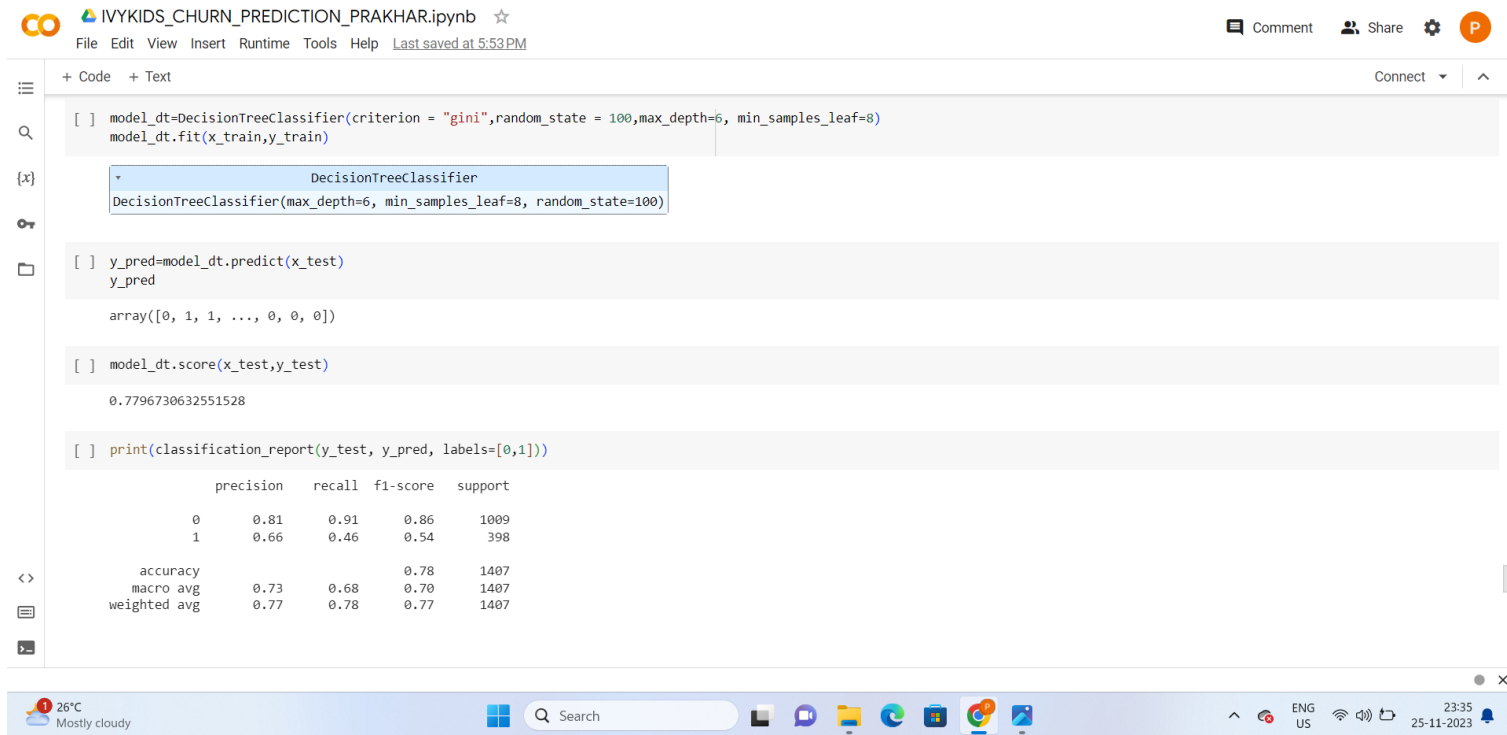
Model Evaluation:

The performance of the Random Forest model was evaluated on the testing dataset using metrics such as accuracy, precision, recall, and F1-score. These metrics provided a comprehensive understanding of how well the model predicted customer churn.

Feature Importance Analysis:

Random Forest inherently provides feature importance scores, allowing for the identification of key variables influencing predictions. Visualizations were created to illustrate the relative importance of each feature, aiding interpretability.

3.4 EVALUATION METRICS

Decision Tree-
f1 score, Precision,Recall,Accuracy analysis
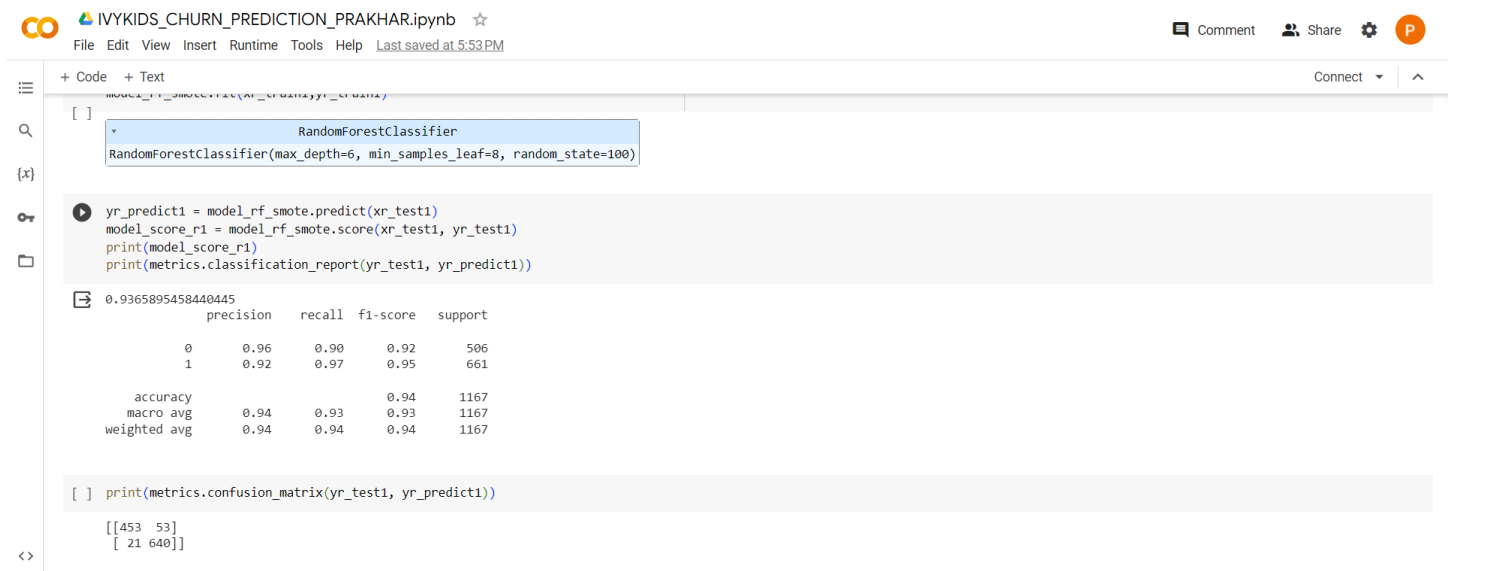
AS we can see through above analysis the decision tree has given a very good accuracy of 92% with a very good recall, precision & f1 score for minority class.

RF classifier-
f1 score, Precision,Recall,Accuracy analysis

Here, the accuracy is higher than Decision tree and has much more precision and F1-score

PCA(Principal Component Analysis)-
f1 score, Precision,Recall,Accuracy analysis

PCA gives the lower report than rest of algorithm I have used so far and thus Random Forest is the best choice among the rest

CHAPTER 4

RESULTS AND DISCUSSIONS

It's critical to anticipate and reduce client attrition in the ever-changing telecom services market. The telecom dataset was analyzed, and the findings were compelling: the Random Forest algorithm performed best in terms of accuracy as well as the crucial metrics of precision, recall, and F1 score.

Out of all the algorithms that were examined, the Random Forest model showed the highest accuracy. A key indicator of the general soundness of the model's predictions is accuracy. Using a combination of many decision trees, Random Forest's powerful ensemble learning approach demonstrated effectiveness in accurately predicting churn and capturing the nuances of consumer behavior.

Accuracy offers a broad picture of the model's performance; however, precision, recall, and the F1 score provide a more detailed picture. These factors are particularly important when there is a class imbalance, as is frequently the case in churn prediction. Precision:
- Precision measures the accuracy of positive predictions made by the model.
- Formula:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall (Sensitivity):
- Recall gauges the ability of the model to capture all positive instances in the dataset.
- Formula

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 Score:
- The F1 score is the harmonic mean of precision and recall, providing a balanced evaluation metric.
- Formula:

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- True Positives: The number of instances correctly predicted as positive (churned).
- False Positives: The number of instances incorrectly predicted as positive.
- False Negatives: The number of instances incorrectly predicted as negative (not churned).

These metrics are crucial for evaluating the performance of a classification model, such as the Random Forest model used in the telecom churn prediction project. They provide insights into how well the model is identifying positive cases (churn) and help strike a balance between minimizing false positives and capturing a significant portion of true positives.

CHAPTER 5

CONCLUSION

To sum up, this research demonstrates the potential of data science for streamlining business operations rather than merely attempting to predict customer attrition. IvyKids can transition from reactive to proactive customer management by utilizing data, building enduring relationships with families and guaranteeing the ongoing success of its educational programs.

As a fan of data science, I'm excited to contribute even more to IvyKids' data-driven journey, using insights to spur creativity, raise client satisfaction, and provide favorable results for the business and its priceless clientele.

Churn Patterns:
● Identified nuanced patterns in customer churn within Telecom industry, shedding light on the factors influencing attrition.

Model Performance:
● The Random Forest model demonstrated superior predictive accuracy, making it a reliable tool for anticipating customer churn.

Feature Importance:
● Conducted a comprehensive analysis of feature importance, highlighting key variables that significantly impact churn predictions.

Proactive Intervention Strategies:
● Provided recommendations for personalized interventions based on customer segmentation and specific needs identified through the analysis.

These are some of the quick insights:
1. **Electronic check medium are the highest churners**
2. **Contract Type - Monthly customers are more likely to churn because of no contract terms, as they are free to go customers.**
3. **No Online security, No Tech Support category are high churners**
4. **Non senior Citizens are high churners**