# Data Scientist Job Change Prediction

By Raghad Althunayan & Shatha Almoteb

# Table of Contents

# 1. Introduction

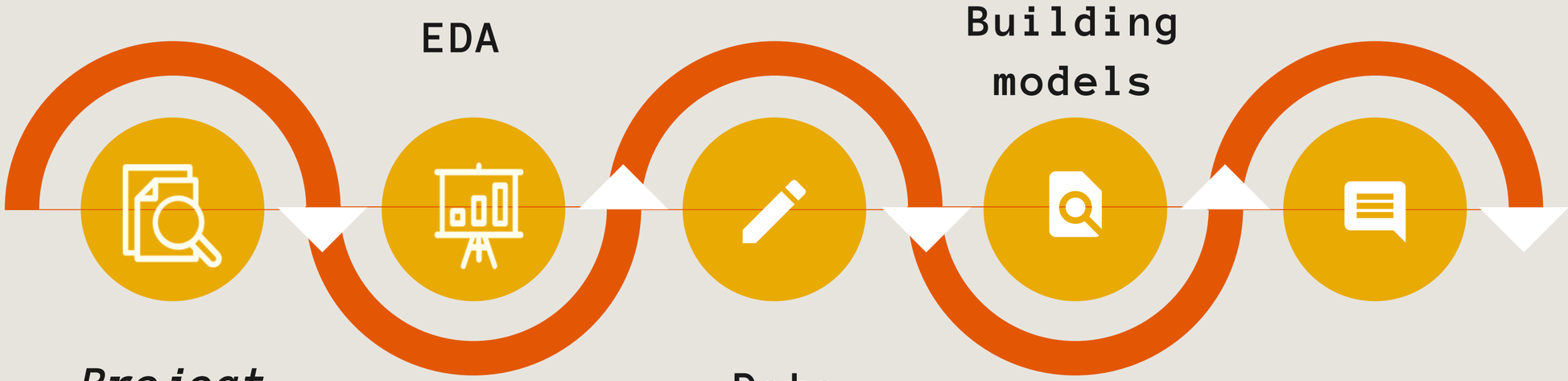
# Introduction

- Will the employee work for the company or look for a new job ?

- It helps to reduce the cost and time as well as the quality of training or planning the courses

# 2. Workflow

# WORKFLOW


EDA
Building models
Project Idea
Data Cleaning & Feature Engineering
Model Deployment

# 3. Data & Design

# Dataset

### Dataset

- From Kaggle

### Size

- 19158 record
- 14  Features

### Target

- Looking for a new job
- Not Looking for a new job

# DATA CLEANING & EDA

Check null.

Deal with outlier.

Remove Unnecessary columns.

Drop duplicates.

Converting categorical values into numeric values

# Data Preparation

## Feature Selection

- Drop 'enrollee_id' and 'city' columns

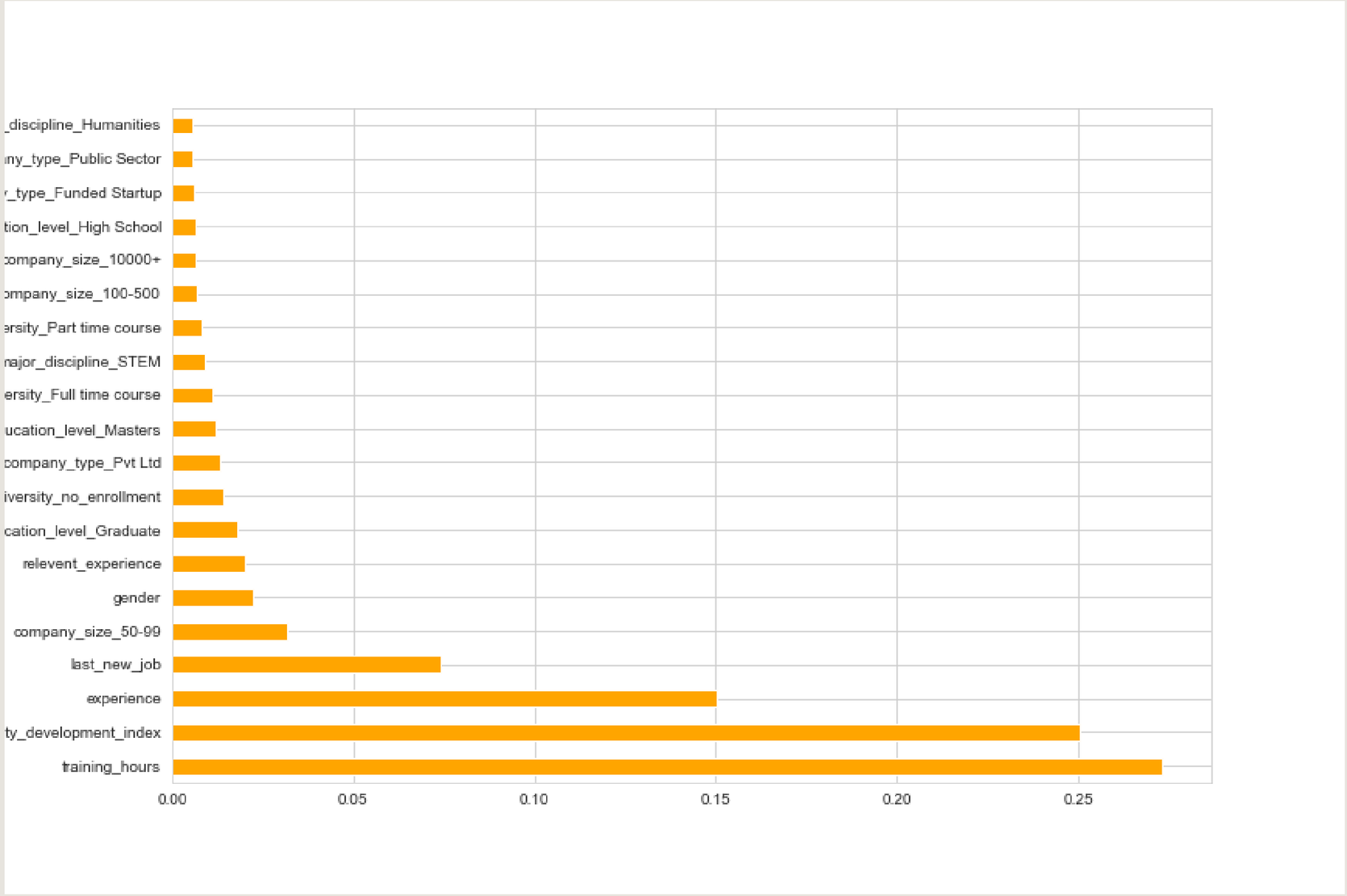
## Feature Enginneering

- label encoding
- get_dummies（Encoding the columns into categorical values.）

## Imbalanced dataset

- SMOTE was use for handling the imbalanced

# Feature importance

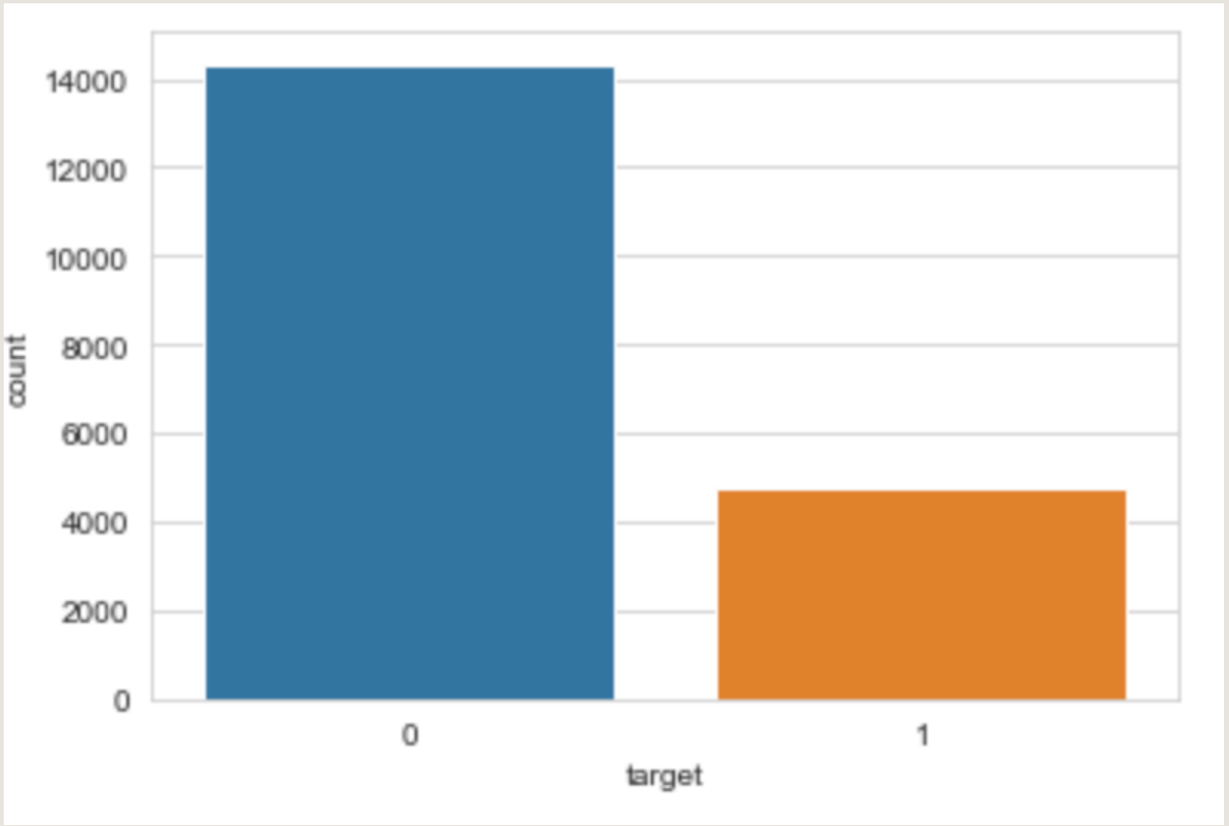



**According to bar chart, these featurs:**

1- training_hours
2- city_development_index
3- experience
4- last_new_job
5-company_size_50-99
6- gender

**Are the most important features .**

# Class Imbalance

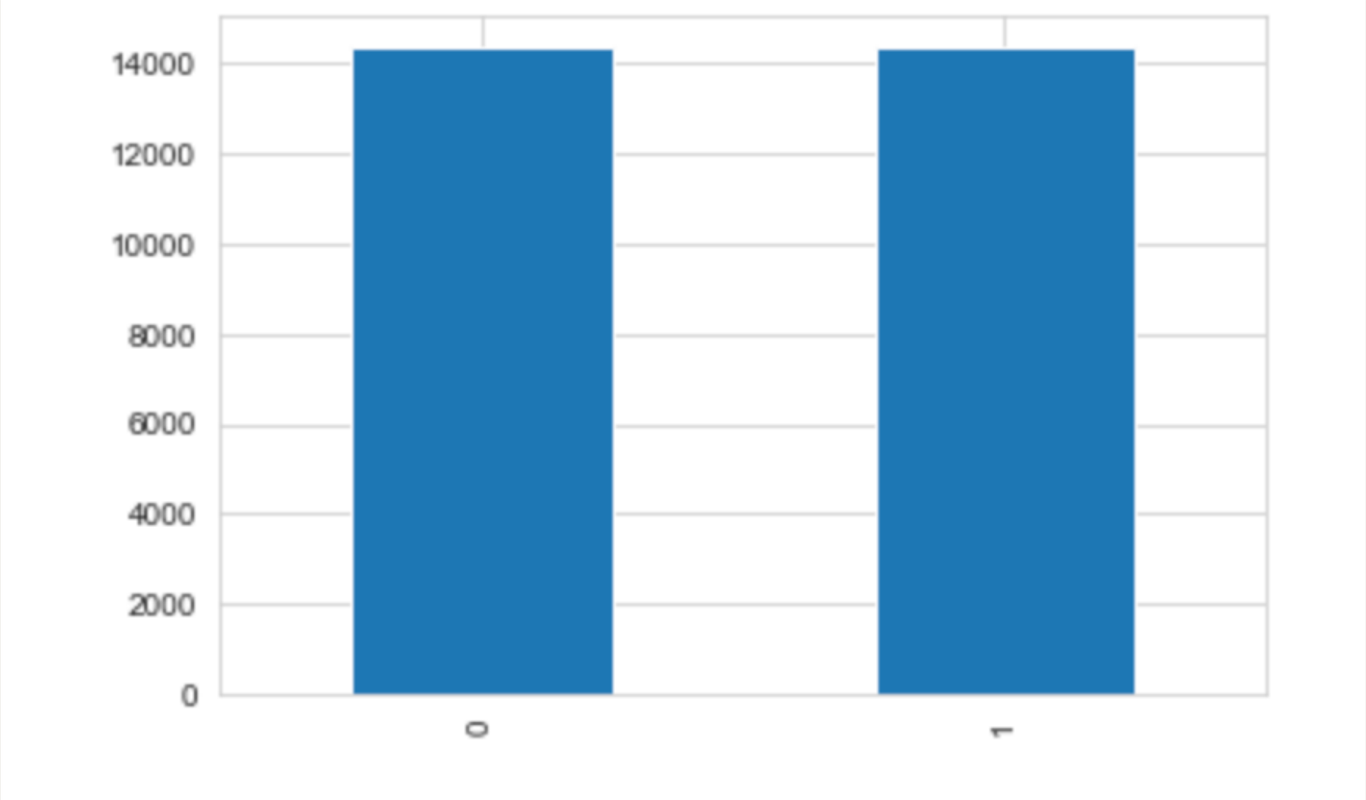



*Target*

0 : 14337

1: 4761

# Solving Class Imbalance




**SMOTE**

0 : 14337

1: 14337

**ADAYSN**

0 : 14337

1: 14337

**Random over sampler**

0 : 14337

1: 14337

# 4.Models

# Classification Models

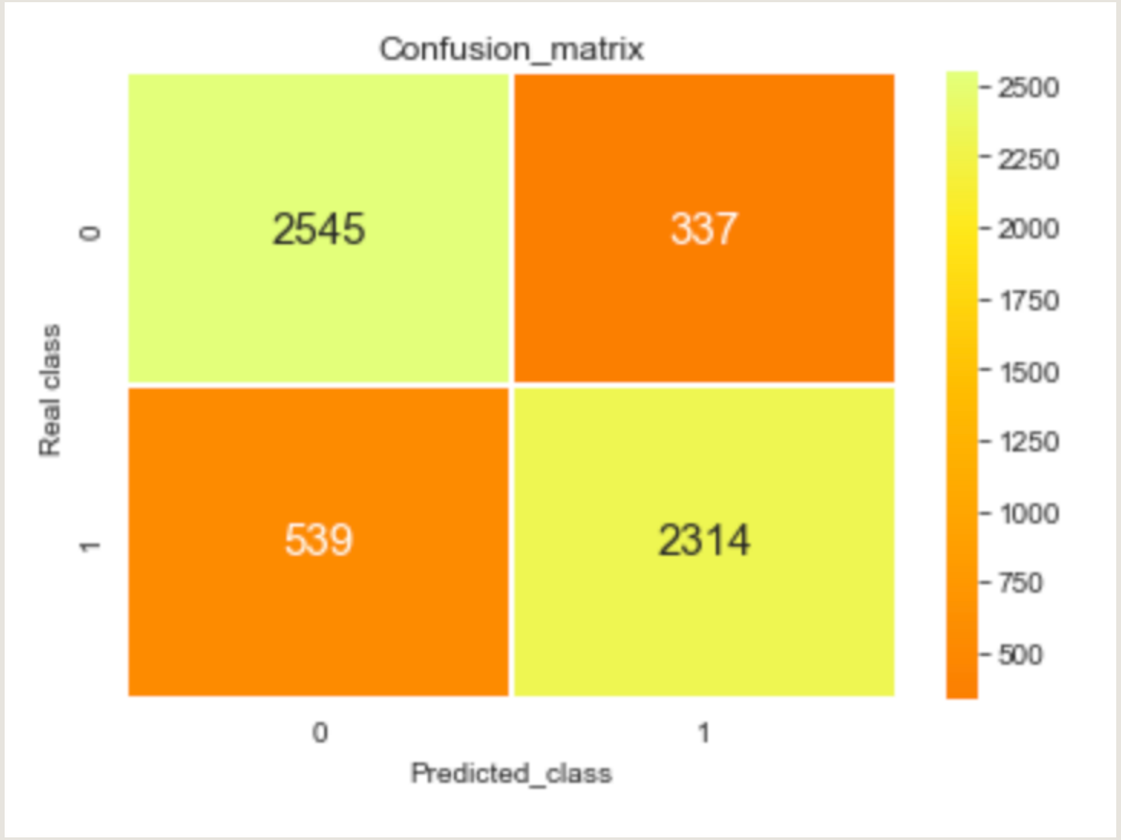
# F1 Score

| | Train | Validation | Test |
|---|---|---|---|
| Baseline Model | 0.77 | 0.77 | - |
| Logistic Regression | 0.77 | 0.77 | 0.77 |
| KNN | 0.99 | 0.75 | 0.76 |
| Decision Tree | 0.81 | 0.80 | 0.80 |
| Random Frost | 0.79 | 0.81 | 0.75 |
| XGBoost | 0.95 | 0.91 | 0.92 |
| SVC | 0.58 | 0.58 | 0.58 |
| GradientBoosting Classifier | 0.84 | 0.83 | 0.83 |
| AdaBoost Classifier | 0.81 | 0.80 | 0.81 |
| MLP Classifier | 0.83 | 0.78 | 0.78 |

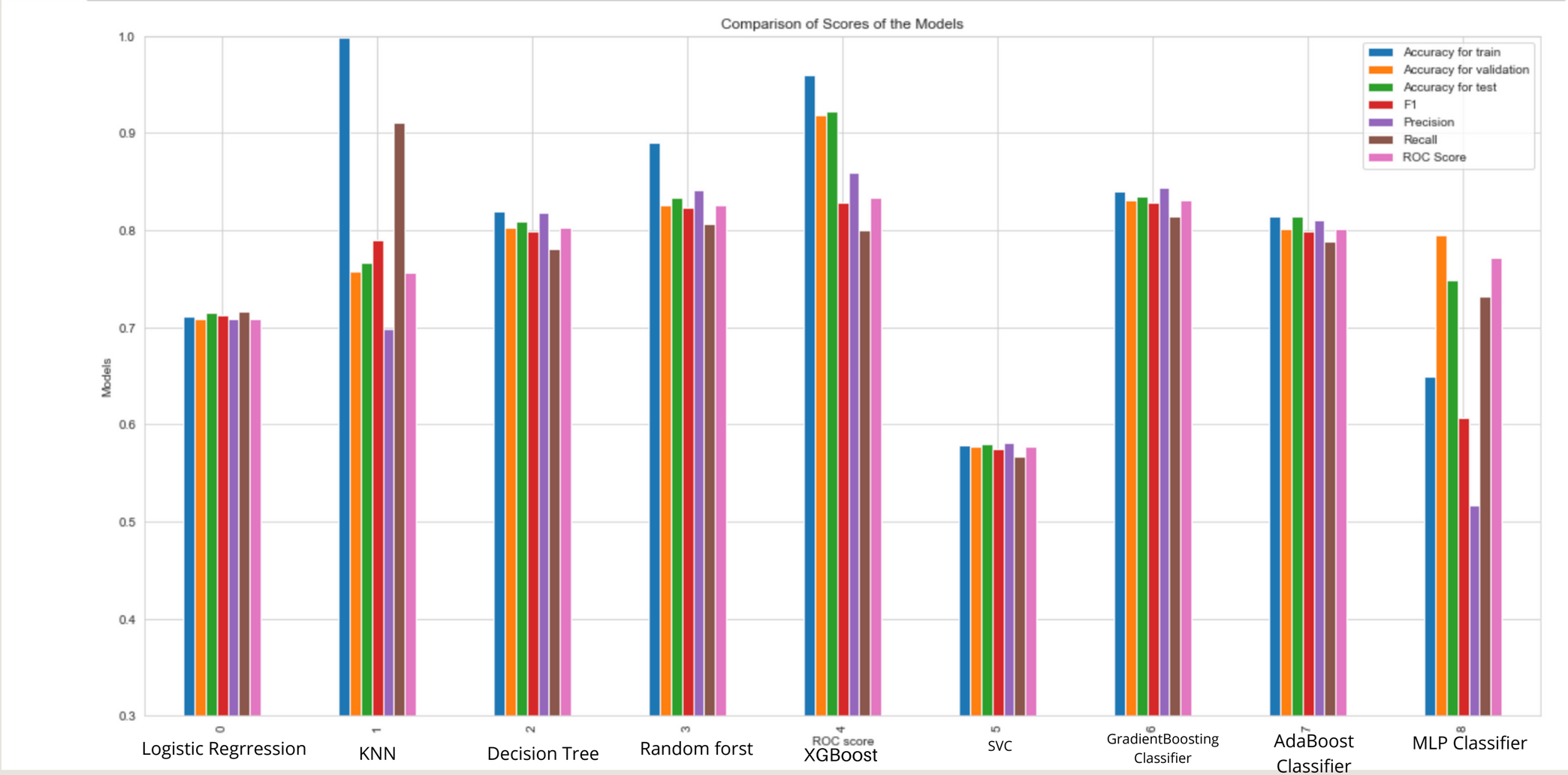
# The Best Model

| | Train | Validation | Test |
|---|---|---|---|
| Baseline Model | 0.77 | 0.77 | - |
| Logistic Regression | 0.77 | 0.77 | 0.77 |
| KNN | 0.99 | 0.75 | 0.76 |
| Decision Tree | 0.81 | 0.80 | 0.80 |
| Random Frost | 0.79 | 0.81 | 0.75 |
| XGBoost | 0.95 | 0.91 | 0.92 |
| SVC | 0.58 | 0.58 | 0.58 |
| GradientBoosting Classifier | 0.84 | 0.83 | 0.83 |
| AdaBoost Classifier | 0.81 | 0.80 | 0.81 |
| MLP Classifier | 0.83 | 0.78 | 0.78 |

# Confusion Matrix and ROC (XGBoost)

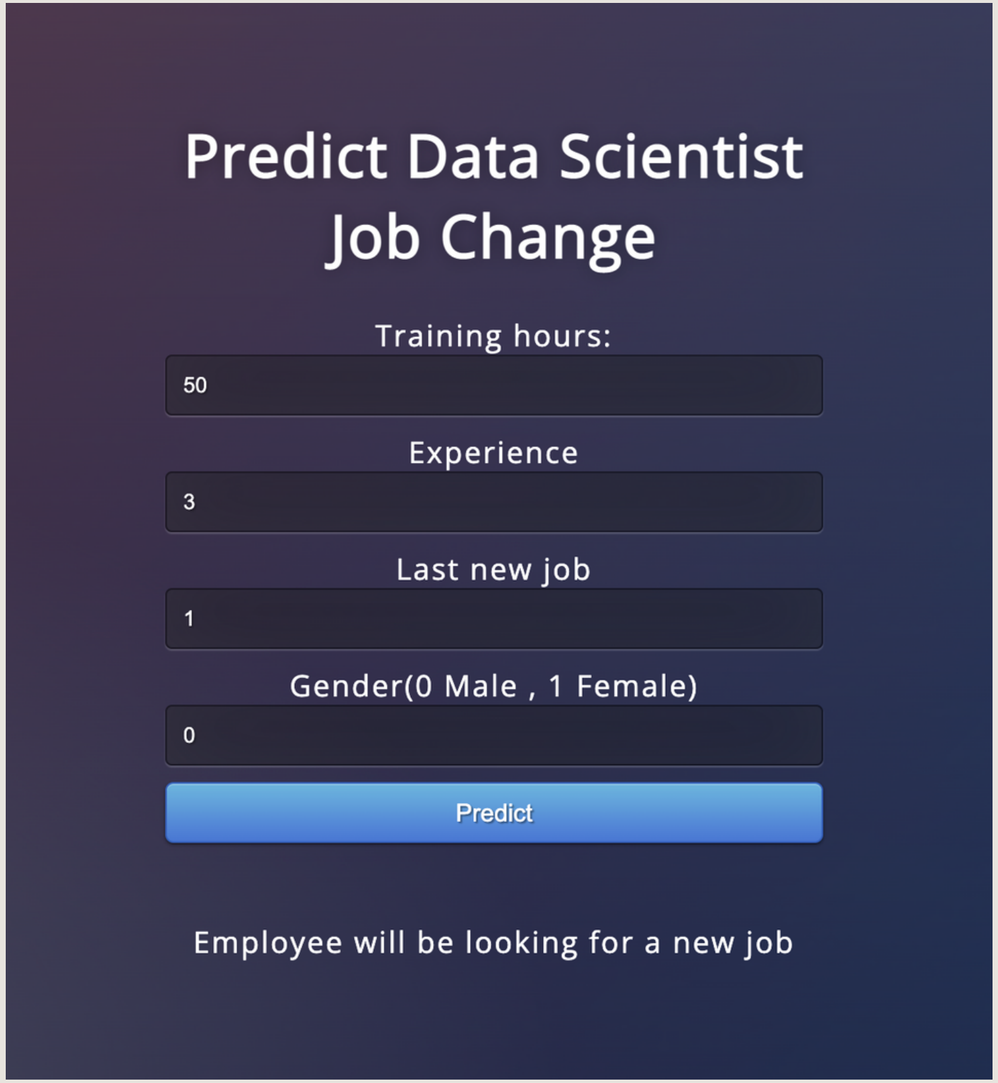



# Comparison Between Models



# 5.Model Deployment

# Model Deployment



# 6.Tools

# Technologies and Libraries


Numpy , Pandas

**Data Cleaning & Manipulation**


Matplotlib , Seaborn

**Visualization**


Sklearn

**Model Building**


Flask

**Web Deployment**

# 7.Conclusion

# Conclusion

**XGBoost provided the best prediction with accuracy score 0.92**

***Future Work;***

- ***Optimizing the model***
- ***Explore additional features***

# Thank you