Job Change of Data Scientists

DATA BOOTCAMP FINAL PROJECT | APRIL 1, 2021

Group Members

Andra Lobo

DATA ENGINEER

Grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.

Brooke Heitshu

DATABASE ADMINISTRATOR

Facilitated and arranged the database environment to support the analytics need of the team working on a project.

Xikang Zhang

DATA SCIENTIST

Facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.

Contents

Group Two

Part 1:

Introduction

JOB CHANGE OF DATA SCIENTISTS

APRIL 1, 2021

The 14 Features of our Dataset

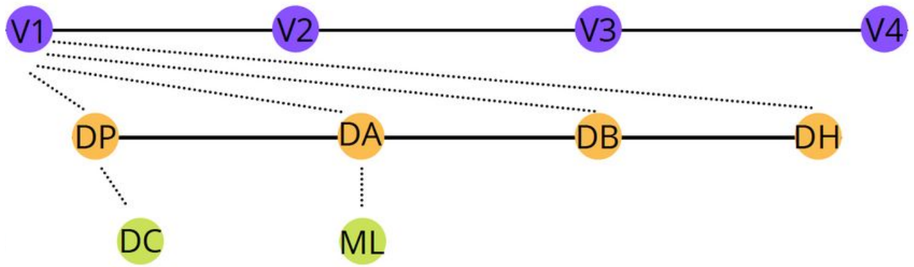
Objectives

What is the likelihood of an employee staying once they complete their training?

What are the key aspects of loyal employees?

Is hiring a less qualified employee more likely to stay after training?

Pipeline



Main

Final Data

Developing

Data Preprocessing, Data Analyzing, Database, Dashboard

Feature

All of our code being worked on is completed in feature branches then pushed to development branches once finished

Part 2:

Database

Preprocessing

- Removed the city & enrollee ID columns
- Converted Categorical values to Numerical
- Created & imported into PostgreSQL database
  1. company_info.csv
  2. personal_info.csv
- Joined the tables into 1

Part 3:

Machine Learning

Machine Learning Process

1 Selected Algorithm

Random Forest Classifier

2 Accuracy

Accuracy score as high as 85%

3 Limitations

Large number of trees that can slow down the algorithm for real time prediction

Results

- After compiling and fitting the training dataset to the model, we have achieved the accuracy scores as 84.6%.

Confusion Matrix

	Predicted staying	Predicted leaving
Actually staying	838	29
Actually leaving	132	51

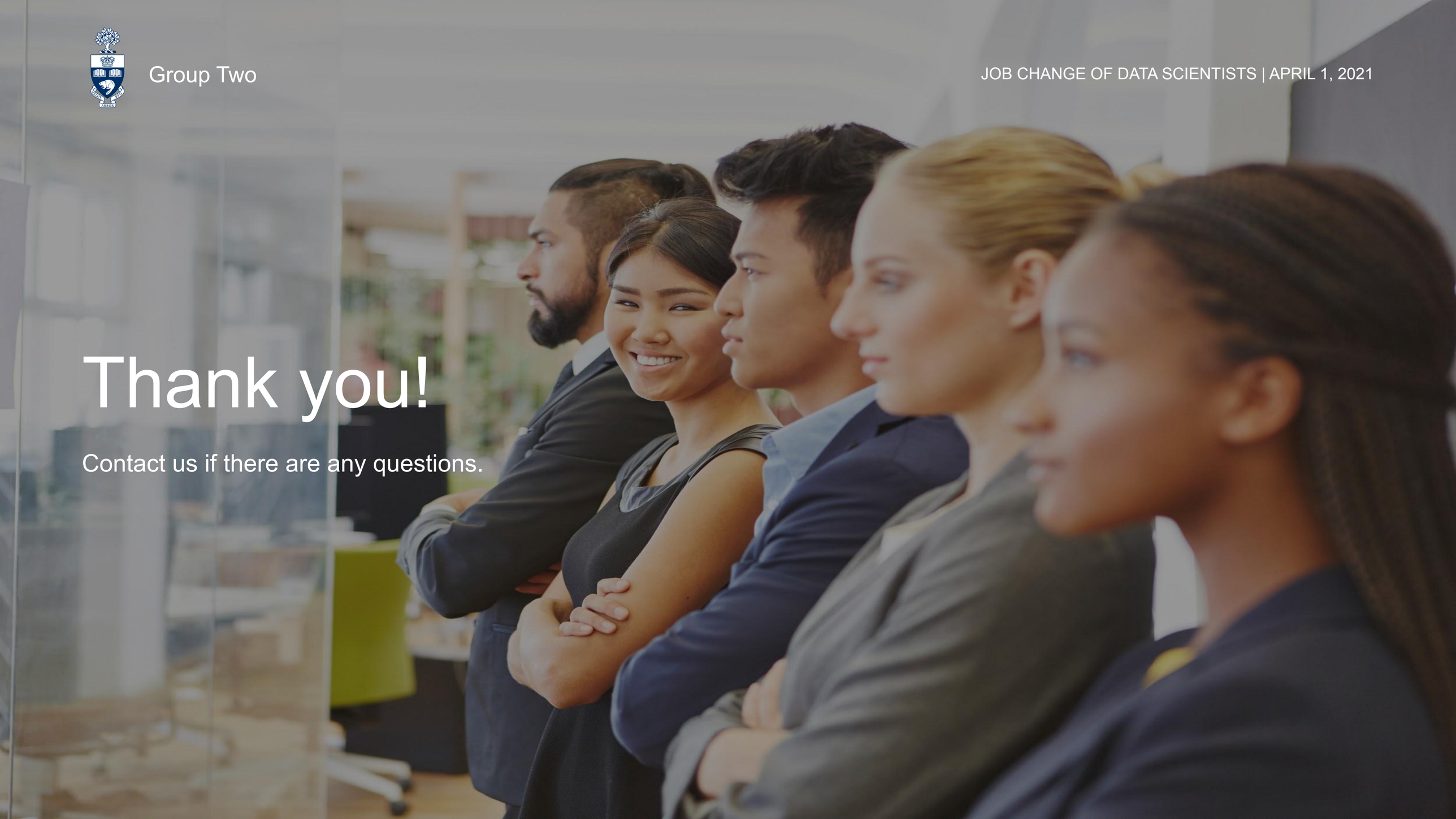
```
Accuracy Score : 0.8466666666666667
Classification Report
              precision    recall  f1-score   support

         0.0       0.86      0.97      0.91       867
         1.0       0.64      0.28      0.39       183

    accuracy                           0.85      1050
   macro avg       0.75      0.62      0.65      1050
weighted avg       0.82      0.85      0.82      1050
```

```
# We can sort the features by their importance.
x=zip(importances,X.columns)
sorted(x,reverse=True)
```

```
[(0.281334865899091, 'city_development_index'),
 (0.1732235693100496, 'training_hours'),
 (0.019148033397059686, 'company_size_50-99'),
 (0.018555184214108754, 'company_size_100-500'),
 (0.017646837475063657, 'experience_>20'),
 (0.017315392321240537, 'company_size_10000+'),
 (0.016784252043938973, 'education_level_Masters'),
 (0.016598347417148263, 'last_new_job_1'),
 (0.01655290415881218, 'education_level_Graduate'),
 (0.015120720947111999, 'company_type_Pvt Ltd'),
 (0.014748307207866485, 'last_new_job_>4'),
 (0.014245123701493289, 'company_size_10/49'),
 (0.014149962795968377, 'company_size_1000-4999'),
 (0.012268219916075077, 'enrolled_university_no_enrollment'),
 (0.012091474150942812, 'last_new_job_2'),
 (0.011355099481823661, 'company_size_500-999'),
 (0.0107756896448112, 'experience_10'),
 (0.010740549573722084, 'gender_Male'),
```



Thank you!

Contact us if there are any questions.