# Final Project
## Data Scientists Recruiting Optimization

### Team 7 - Five Guys

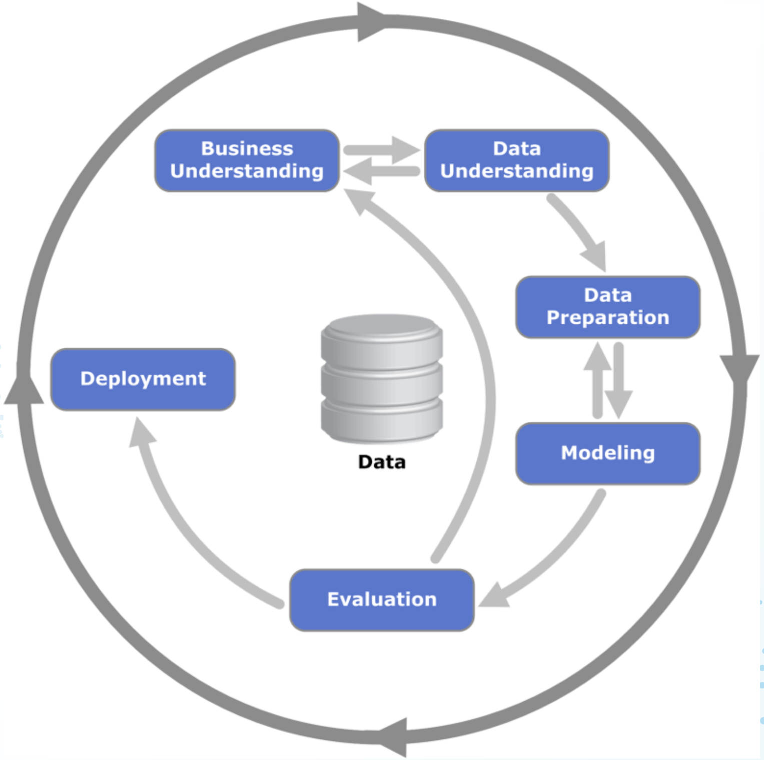
Will Jiang, Maggie Meng, Paakhi Srivastava, Yulin Gai, Joe Ebby Karuthedath

# Review Key Points of Classification

- Definition of **Classification Task**
  - A supervised learning technique to predict whether an individual belongs to a certain categorical class

- Examples of **Classification Algorithms**
  - Decision Tree
  - K-Nearest Neighbors (KNN)
  - Logistic Regression

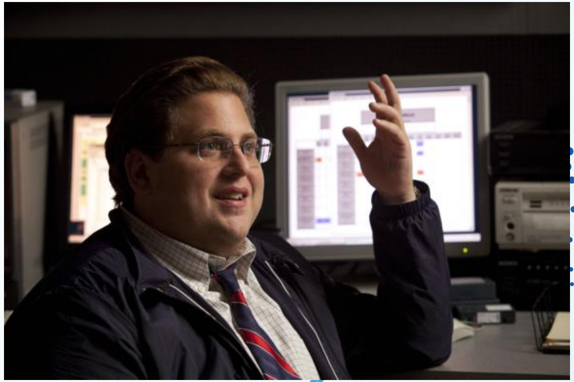
# Review of the Data Mining Process: CRISP



# 1.Business Understanding

# Business Understanding

- US companies spent more than **$70 billion** on training employees last year
- People with good technical skills are **hard to find** and harder to retain
- Our company is training people to fill up their Data Science Job vacancies
- But **not** all people who enroll in the training are **really looking for a job change**!
- This leads to a **loss of** the companies valuable **resources**



# Business Understanding: Predict "Who takes up the new JOB?"

- Predicting who takes up the new job with our company to fill up the data science vacancies can lead to significant savings for the company.
- **OBJECTIVE** : Create a classification model to predict if a person is a potential employee or not.
  - This can help the company to target potential candidates to train
  - This targeted approach means that more of the trainees fill up the vacancies.

# 2.Data Understanding

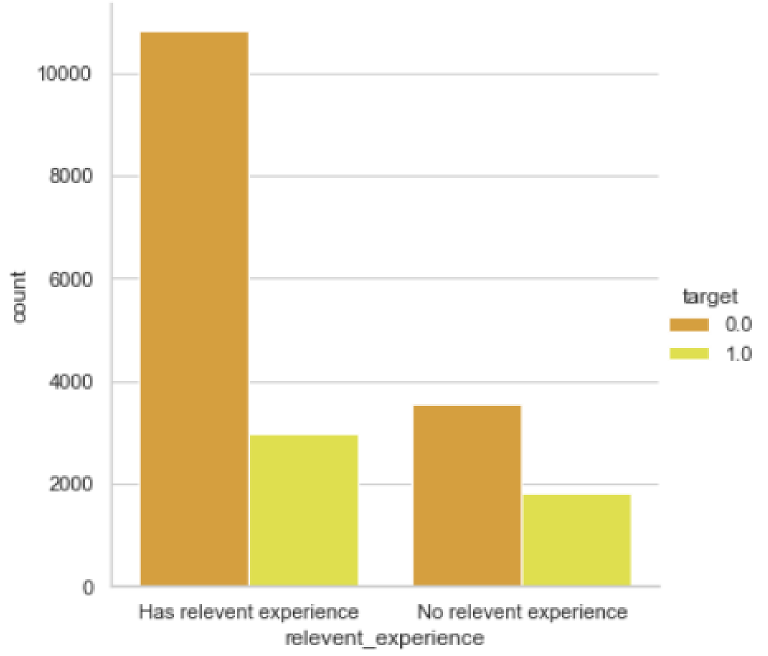
# Data Understanding: Dataset

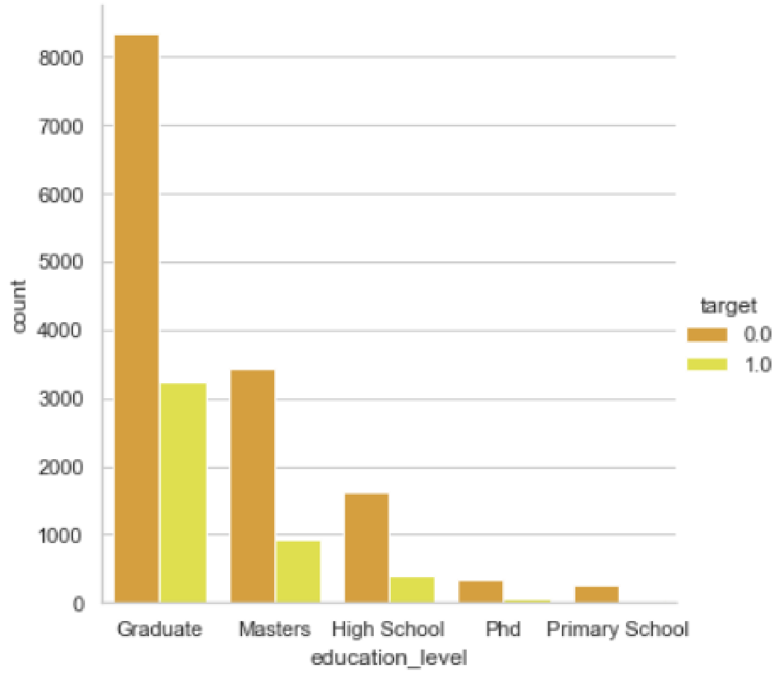
- HR Analytics: Job Change of Data Scientists
  - Characteristics of Data Scientists who received training for a job change
  - 19158 Observations
  - 13 Features
    - Mostly categorical: gender, relevant experience, enrolled in university, education level, major discipline, company size and type, etc.
    - Some Numerical: training hours, last new job, city development index

# Data Understanding: Exploratory Data Analysis

- Complete EDA on the dataset
  - Summary statistics
  - Missing values
  - Bar Charts for categorical variables
  - Examination of correlation

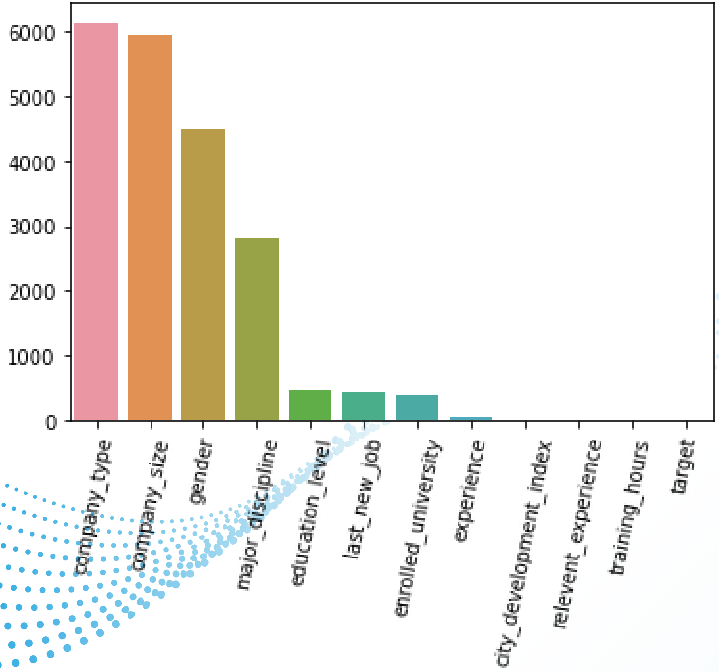
# Exploratory Data Analysis



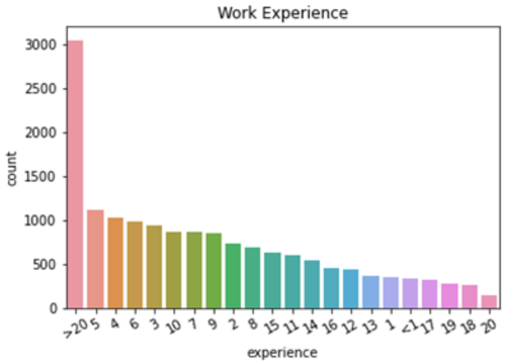
# 3.Data Preparation

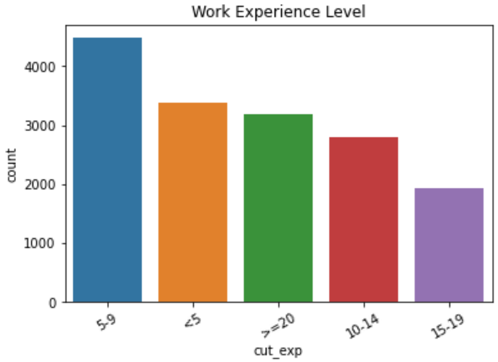
# Data Preparation: Missing Value




- **Drop the missing value** in “University”, “Education Level”, “Major Discipline”, “Experience” and “Last Job”.
- **Replace missing value** in “Gender” with “Other”.
- **Replace missing value** in “Company Size” and “Type” with mode of those attributes.

# Data Preparation:
# Feature Engineering and Selection

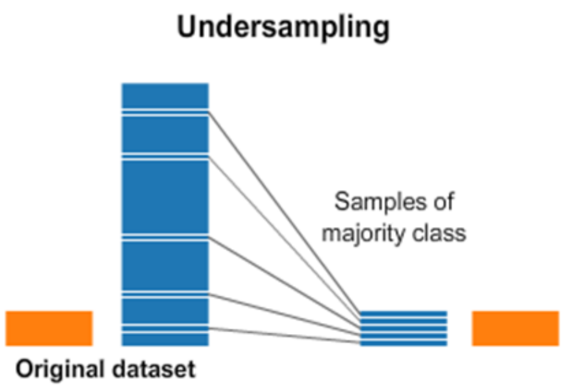


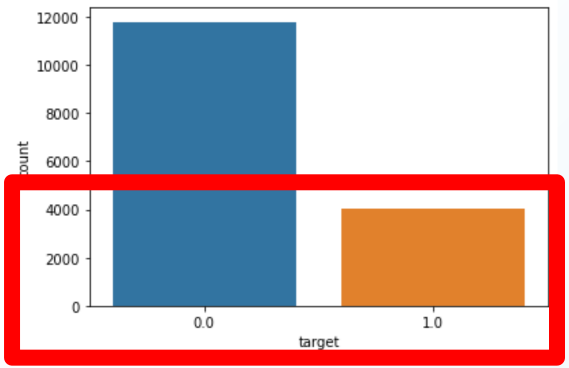




- Sort and segment "Work Experience" into 5 bins:
  - Less than 5 years, 10 to 14 years, 15 to 19 years, more than 20 years working experience

- Finally, our features include:
  - Numeric: CityDevelopmentIndex, TrainingHour (2)
  - Categorical: Gender, RelevantExperience, University, Major, CompanySize, CompanyType, LastJobs, WorkExperience (8)

# Data Preparation: Data Balancing




- **Undersampling**
  - Randomly sample majority instances and repeat until the dataset contains an equal number of each class
  - Final dataset includes more than 8,000 data points, with equal size of both targets

# 4.Modeling

# Modeling

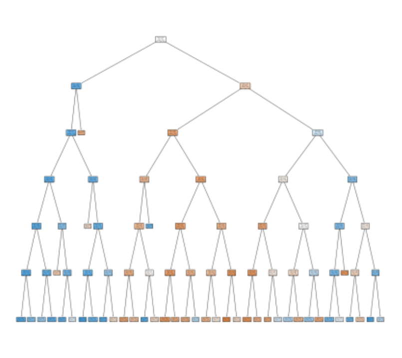
- Try out three different **models** to determine the best performance one:
  - Logistic Regression, KNN, Decision Tree
- Utilize a grid search for **hyperparameter tuning**
- Use **nested cross-validation** to evaluate generalization performance
  - 5 folds in the inner and outer loops
  - Accuracy scoring metric
- Conduct additional data preparation for the KNN and Logistic Regression model:
  - Use pipeline
  - Standardize numeric features by using training data

# Model Comparison

| | Decision Tree | Logistic Regression | KNN |
|---|---|---|---|
| **pros** | - Easy to implement<br>- Computational cheap<br>- Model **comprehensibility** | - Maximum control<br>- **Fast scoring**<br>- Robustness | - **“Lazy”** model<br>- Easy to implement and use<br>- Robustness<br>- **No** statistical / distribution **assumption** required |
| **cons** | Tend to **overfitting** | **Not flexible** for larger training set | - Take **more time** to perform estimation<br>- Requires a lot of storage<br>- Lack of interpretable model<br>- Curse of dimensionality |

# Model Results

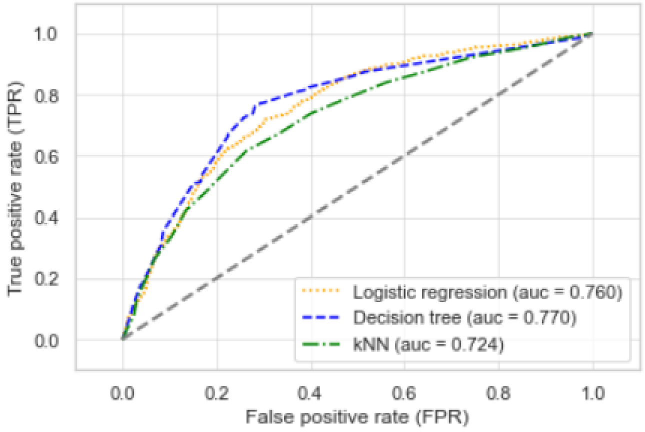
| | Decision Tree | Logistic Regression | KNN |
|---|---|---|---|
| Hyperparameter Grid Search | criterion: gini, max_depth: 6, min_samples_leaf: 7 | C: 0.01, penalty: L2 | n_neighbors: 27, weights: uniform |
| Non-nested CV accuracy | **0.744** ★ | 0.705 | 0.681 |



# 5.Evaluation

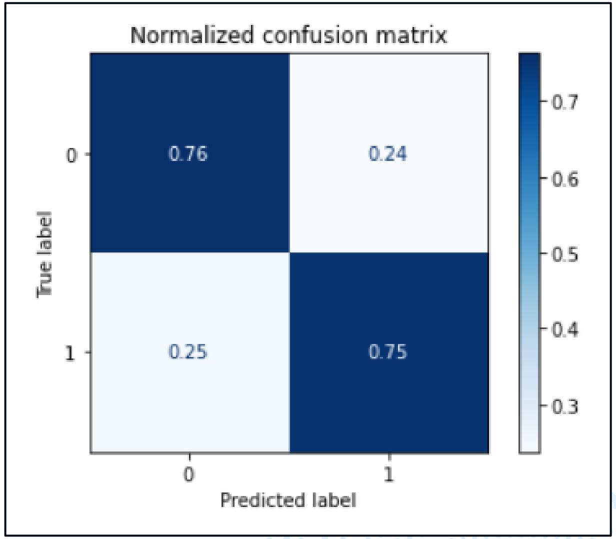
# Result Evaluation

| Model Classifier Type | Accuracy | AUC |
|---|---|---|
| K-Nearest Neighbor | 0.681 | 0.73 |
| Logistic Regression | 0.705 | 0.76 |
| Decision Tree ★ | 0.744 | 0.78 |




- Decision Tree model has the highest accuracy score and AUC estimator → Decision Tree is the best performing model
- The higher the **AUC**, the better the model is at **distinguishing whether the employee will leave after the training session**
- Since we have plenty of dummy variables in our data set, and **decision trees can effectively handle non-linear data sets**, our data mining result is logically feasible

# ROI




Cost/Benefit Information

| | p | n |
|---|---|---|
| Y | $198.8k | -$1.2k |
| N | 0 | 0 |

Sources: https://elmlearning.com/how-much-does-employee-training-really-cost/
https://smallbusinessmattersonline.com/revenue-per-employee-calculation/

- We can use our Decision Tree expected rates matrix and cost/benefit matrix to calculate our **expected value**.
- We estimate the cost/benefit information from online sources. (Average revenue per employee = $200k)
- After multiplying two matrices and summing up all the elements, we can get our expected profit, which is around **$151k**.

# 6.Deployment

# Deployment

- The predictive model will determine if the candidate will take up the new job with the company to fill up the data science roll or not
- Ethical Considerations- Gender Bias! The data has a bias towards the male candidates, which is reflected in the model. This should be taken into consideration by the HR manager while hiring.
- Risks:
    - It is possible that the model can miss impressive candidates
- The company can use this model to
    - Target potential candidates to train
        - Save significant money spent on training
        - Filter the incoming applications
        - Use the model as a reference, to build more models to fill other positions

# Prediction: Will Jarvis leave?

## Jarvis

*MSBA candidate*

Gender: Male

Education_level: Masters

Major_discipline: STEM

City_development_index: 0.9

Relevent_experience: 0

Experience: 0

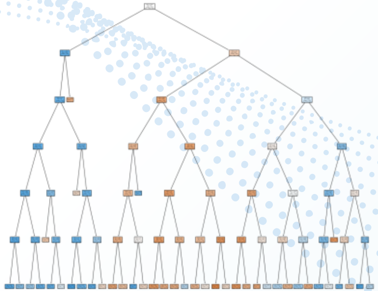
Last_new_job: 0

Company_size: No

Company_type: No

Training_hours: 65(avg)

Hire me!



**Decision tree model training**

# 0

## Not looking for job change!

NO!

# Thank You!

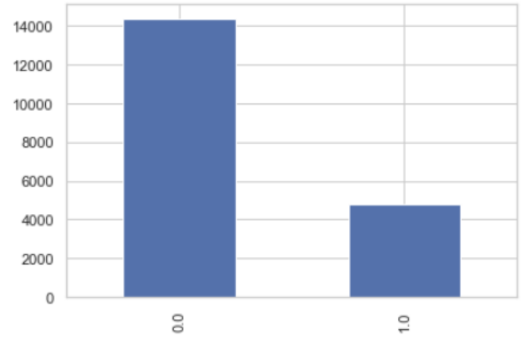
## Questions?

# Appendix

# EDA

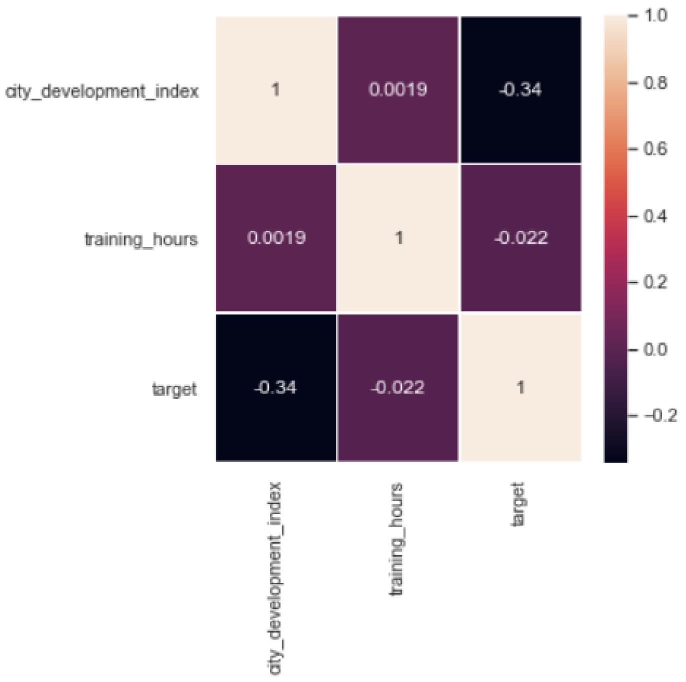
```
df['target'].value_counts().plot(kind='bar')
```

<AxesSubplot:>




```
df.isna().sum()
```

```
city_development_index       0
gender                    4508
relevent_experience          0
enrolled_university        386
education_level            460
major_discipline          2813
experience                  65
company_size              5938
company_type              6140
last_new_job               423
training_hours               0
target                       0
dtype: int64
```



# EDA Continued