# E-RECRUITMENT – HOW IT CONTRIBUTES TO GENDER INEQUALITY

Presented by:

Yifan CAO and Toni LAW

# Content

# Introduction

**Previous literature have focused on:**

- **Efficiency and accuracy of hiring process**
- **The perspective of recruiters**

**Our research aims to:**

- **To identify bias (gender bias)**
- **To improve in reliability**

# Theory & Hypothesis (Expectation of theory)

**Theory:**

Gender bias and gender inequality in the workplace is positively correlated

**Hypothesis:**

Gender inequality in job attainment is associated with gender bias during the E-recruitment process

# Related work- Inequality in employment

**Gender bias in employment:**

Investigate 14 cities 316 law firm in America
To find out possibility of hiring in elite labor market:
higher class men > higher class women

**Hierarchy bias in employment:**

A qualitative case study of how elite reproduction occurs in hiring.

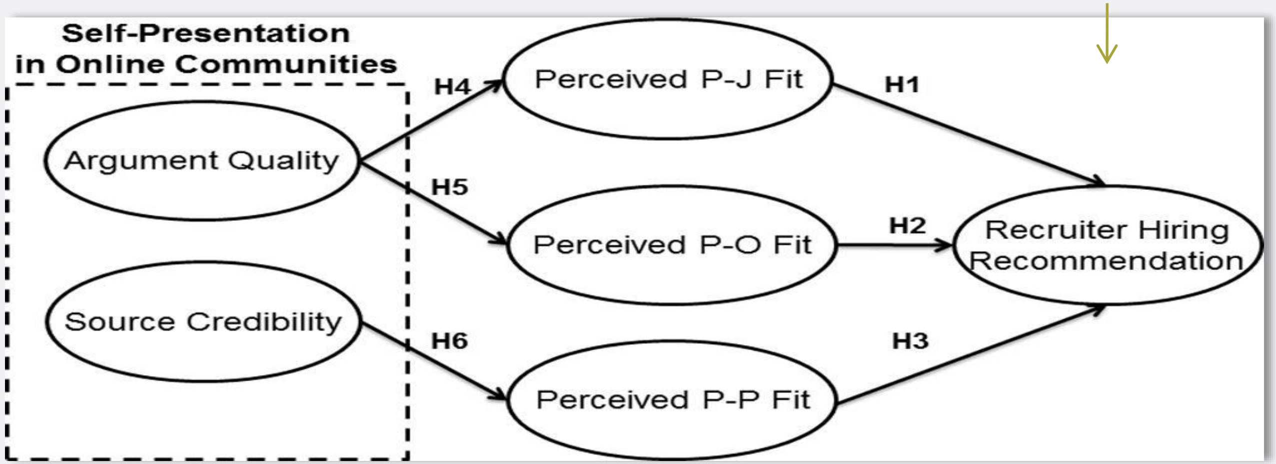
From 2006 to 2008, 120 semistructured interviews (40 interviews per firm type) with professionals involved in undergraduate and graduate hiring in
top-tier consulting, banking, and law firms.

"I originally intended to study gender in hiring. After the initial coding rounds, however, it
became clear that culture and socioeconomic status were highly salient bases of evaluation and stratification in these firms."

Lauren A. Rivera, & András *Tilcsik. (201*6). Class Advantage, Commitment Penalty: The Gendered Effect of Social *Class Signals in an Elite* Labor Market. American Sociological Review, 81(6), 1097-1131.
Rivera, L. A. (2015). Pedigree. Princeton: Princeton University Press.

# Related work - Finding features of  E-recruitment

Self-presentation → Matching mechanism → Hiring-recommendations




Methods:
**focus group, interview, questionnaire, participant observation, statistical significance**

Features & Hypothesis:

(1) portrait
(2) profile summary
(3) experience
(4) volunteer experience & causes
(5) projects
(6) languages
(7) certifications
(8) publications
(9) education
(10)discussion posts and comments
(11) recommendations
(12) endorsed skills and expertise
(13) interests
(14) honours and awards.

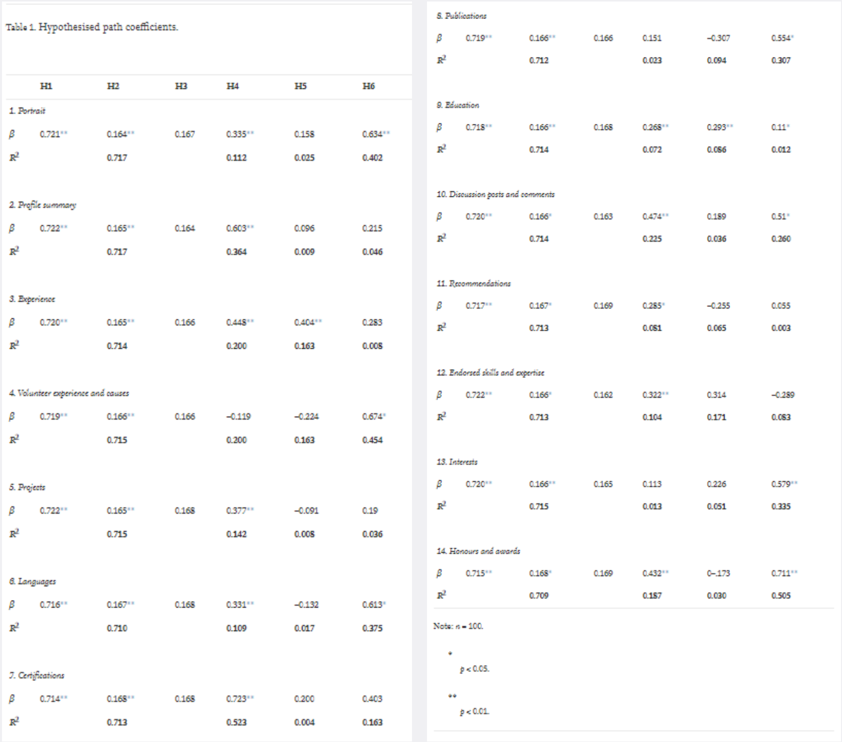


Table 1. Hypothesised path coefficients.

| | H1 | H2 | H3 | H4 | H5 | H6 |
|---|---|---|---|---|---|---|
| *1. Portrait* | | | | | | |
| $\beta$ | 0.721** | 0.164** | 0.167 | 0.335** | 0.158 | 0.634** |
| $R^2$ | | 0.717 | | 0.112 | 0.025 | 0.402 |
| *2. Profile summary* | | | | | | |
| $\beta$ | 0.722** | 0.165** | 0.164 | 0.603** | 0.096 | 0.215 |
| $R^2$ | | 0.717 | | 0.364 | 0.009 | 0.046 |
| *3. Experience* | | | | | | |
| $\beta$ | 0.720** | 0.165** | 0.166 | 0.448** | 0.404** | 0.283 |
| $R^2$ | | 0.714 | | 0.200 | 0.163 | 0.008 |
| *4. Volunteer experience and causes* | | | | | | |
| $\beta$ | 0.719** | 0.166** | 0.166 | −0.119 | −0.224 | 0.674* |
| $R^2$ | | 0.715 | | 0.200 | 0.163 | 0.454 |
| *5. Projects* | | | | | | |
| $\beta$ | 0.722** | 0.165** | 0.168 | 0.377** | −0.091 | 0.19 |
| $R^2$ | | 0.715 | | 0.142 | 0.008 | 0.036 |
| *6. Languages* | | | | | | |
| $\beta$ | 0.716** | 0.167** | 0.168 | 0.331** | −0.132 | 0.613* |
| $R^2$ | | 0.710 | | 0.109 | 0.017 | 0.375 |
| *7. Certifications* | | | | | | |
| $\beta$ | 0.714** | 0.168** | 0.168 | 0.723** | 0.200 | 0.403 |
| $R^2$ | | 0.713 | | 0.523 | 0.004 | 0.163 |
| *8. Publications* | | | | | | |
| $\beta$ | 0.719** | 0.166** | 0.166 | 0.151 | −0.307 | 0.554* |
| $R^2$ | | 0.712 | | 0.023 | 0.094 | 0.307 |
| *9. Education* | | | | | | |
| $\beta$ | 0.718** | 0.166** | 0.168 | 0.268** | 0.293** | 0.11* |
| $R^2$ | | 0.714 | | 0.072 | 0.086 | 0.012 |
| *10. Discussion posts and comments* | | | | | | |
| $\beta$ | 0.720** | 0.166* | 0.163 | 0.474** | 0.189 | 0.51* |
| $R^2$ | | 0.714 | | 0.225 | 0.036 | 0.260 |
| *11. Recommendations* | | | | | | |
| $\beta$ | 0.717** | 0.167* | 0.169 | 0.285* | −0.255 | 0.055 |
| $R^2$ | | 0.713 | | 0.081 | 0.065 | 0.003 |
| *12. Endorsed skills and expertise* | | | | | | |
| $\beta$ | 0.722** | 0.166* | 0.162 | 0.322** | 0.314 | −0.289 |
| $R^2$ | | 0.713 | | 0.104 | 0.171 | 0.083 |
| *13. Interests* | | | | | | |
| $\beta$ | 0.720** | 0.166** | 0.165 | 0.113 | 0.226 | 0.579** |
| $R^2$ | | 0.715 | | 0.013 | 0.051 | 0.335 |
| *14. Honours and awards* | | | | | | |
| $\beta$ | 0.715** | 0.168* | 0.169 | 0.432** | 0–.173 | 0.711** |
| $R^2$ | | 0.709 | | 0.187 | 0.030 | 0.505 |

Note: $n = 100$.

* $p < 0.05$.

** $p < 0.01$.


Chiang, Johannes Kuo-Huie, & Suen, Hung-Yue. (2015). Self-presentation and hiring recommendations in online communities: Lessons from LinkedIn. *Computers in Human Behavior*, 48, 516-524.
Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. *The International Journal of Human Resource Management*, 28(1), 3–26. doi:10.1080/09585192.2016.1244699
Jansen, Anne, König, Cornelius J, Stadelmann, Eveline H, & Kleinmann, Martin. (2012). Applicants' Self-Presentational Behavior: What Do Recruiters Expect and What Do They Get? Journal of Personnel Psychology, 11(2), 77-85.

# Related works - Computational method

## [Big data contributions to human resource management: a systematic review](https://doi.org/10.1080/09585192.2019.1674357)

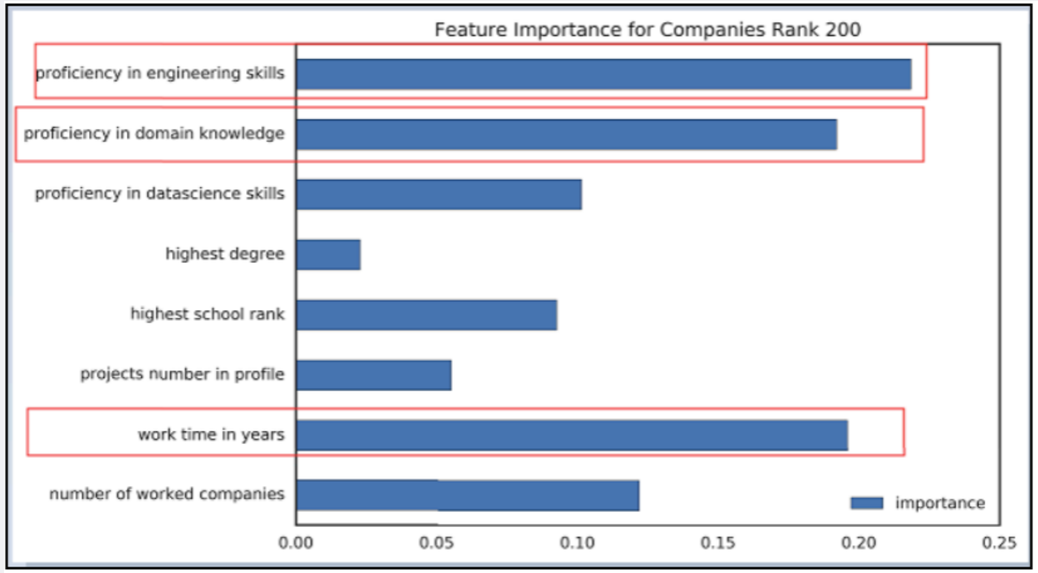
**Table 2.** Distribution of documents per year and cluster. (Table view)

| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|---|---|---|
| HR research and practice | | | | 3 | 1 | 1 | 5 |
| Selecting, hiring | | | 3 | 1 | | | 4 |
| Assessment, development | | | | 1 | 2 | 1 | 4 |
| Information, learning, knowledge | 2 | 2 | 5 | 2 | 5 | 3 | 19 |
| Strategic, efficiency, performance | 1 | 1 | 2 | 2 | 2 | 1 | 9 |
| Total | 4 | 3 | 10 | 9 | 10 | 6 | 41 |

**e.g. 1）Predicting the Probability and Salary to Get Data Science Job in Top Companies by Gradient Boosting Decision Tree (GBDT algorithm)**

Data features from the LinkedIn profile of candidate (above 18 features):

- Username: name of the candidate which will not use in this study
- Summary: self-written personal summary
- Education: University name, degree, and campus performance summary
- Skillsets: names of skills with endorsement value from first-degree connections
- Organizations: professional and academic organizations
- Publications and patents: intellectual property record
- Recommendations: comments on historical performance from very close first degree connections
- Certification and courses: record of training and coursework history
- Current job: company name, job title, responsibility, location, occupation time
- Past jobs: company name, job title, responsibility, location, occupation time





Marivate, V., & Moorosi, N. (2017). Employment relations: A data driven analysis of job markets using online job boards and online professional networks. *Proceedings of the International Conference on Web Intelligence*. doi:10.1145/3106426.3115589

Garcia-Arroyo, J., & Osca, A. (2019). Big data contributions to human resource management: a systematic review. *The International Journal of Human Resource Management*, 1–26. https://doi.org/10.1080/09585192.2019.1674357

Hazra, S., & Sanyal, S. (2016). Recruitment Prediction Using Id3 Decision Tree. *International Journal of Advance Engineering and Research Development*, 3(10). doi:10.21090/ijaerd.031010

Situ, W., Zheng, L., Yu, X., & Daneshmand, M. (2017). Predicting the Probability and Salary to Get Data Science Job in Top Companies. *Industrial and Systems Engineering Research Conference*.

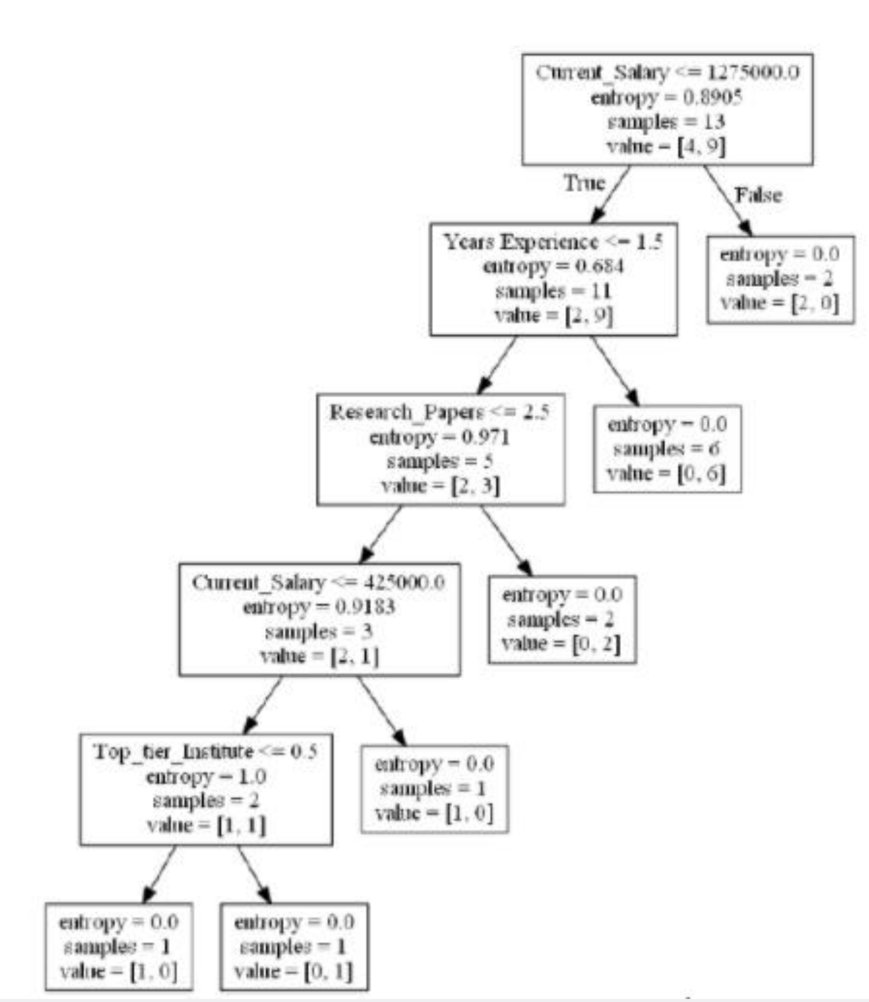
# Related works - Computational method

## e.g. 2）Recruitment Prediction Using ID3 Decision Tree

TABLE 1

Attributes and their possible values

| PARAMETERS | DESCRIPTION | POSSIBLE VALUES |
|---|---|---|
| Years of experience | Shows the number of years of experience of the candidate. | Numeric ( $\geq$ 0) |
| Employed | Checks if the candidate is currently employed. | { Y (Yes), N (No) } |
| Previous employers | Lists the number of previous employers. | Numeric ( $\geq$ 0) |
| Level of education | Shows the highest level of degree pursued by the candidate. | { BS (Bachelors), MS (Masters), PhD (Post graduate) } |
| Top tier institute | Checks if the candidate pursued his/her education from a reputed institution. | { Y (Yes), N (No) } |
| Research papers | Lists the number of research papers published by the candidate. | Numeric ( $\geq$ 0) |
| Internships | Shows the number of internships done by the candidate. | Numeric ( $\geq$ 0) |
| Current salary | The salary which the candidate has been receiving on his/her current job. | Numeric ( $\geq$ 0) |




Marivate, V., & Moorosi, N. (2017). Employment relations: A data driven analysis of job markets using online job boards and online professional networks. *Proceedings of the International Conference on Web Intelligence*. doi:10.1145/3106426.3115589

Garcia-Arroyo, J., & Osca, A. (2019). Big data contributions to human resource management: a systematic review. *The International Journal of Human Resource Management*, 1–26. https://doi.org/10.1080/09585192.2019.1674357

Hazra, S., & Sanyal, S. (2016). Recruitment Prediction Using Id3 Decision Tree. *International Journal of Advance Engineering and Research Development*, 3(10). doi:10.21090/ijaerd.031010

Situ, W., Zheng, L., Yu, X., & Daneshmand, M. (2017). Predicting the Probability and Salary to Get Data Science Job in Top Companies. *Industrial and Systems Engineering Research Conference*.

# Gap and framework

Current computational method of E-recruitment study always:

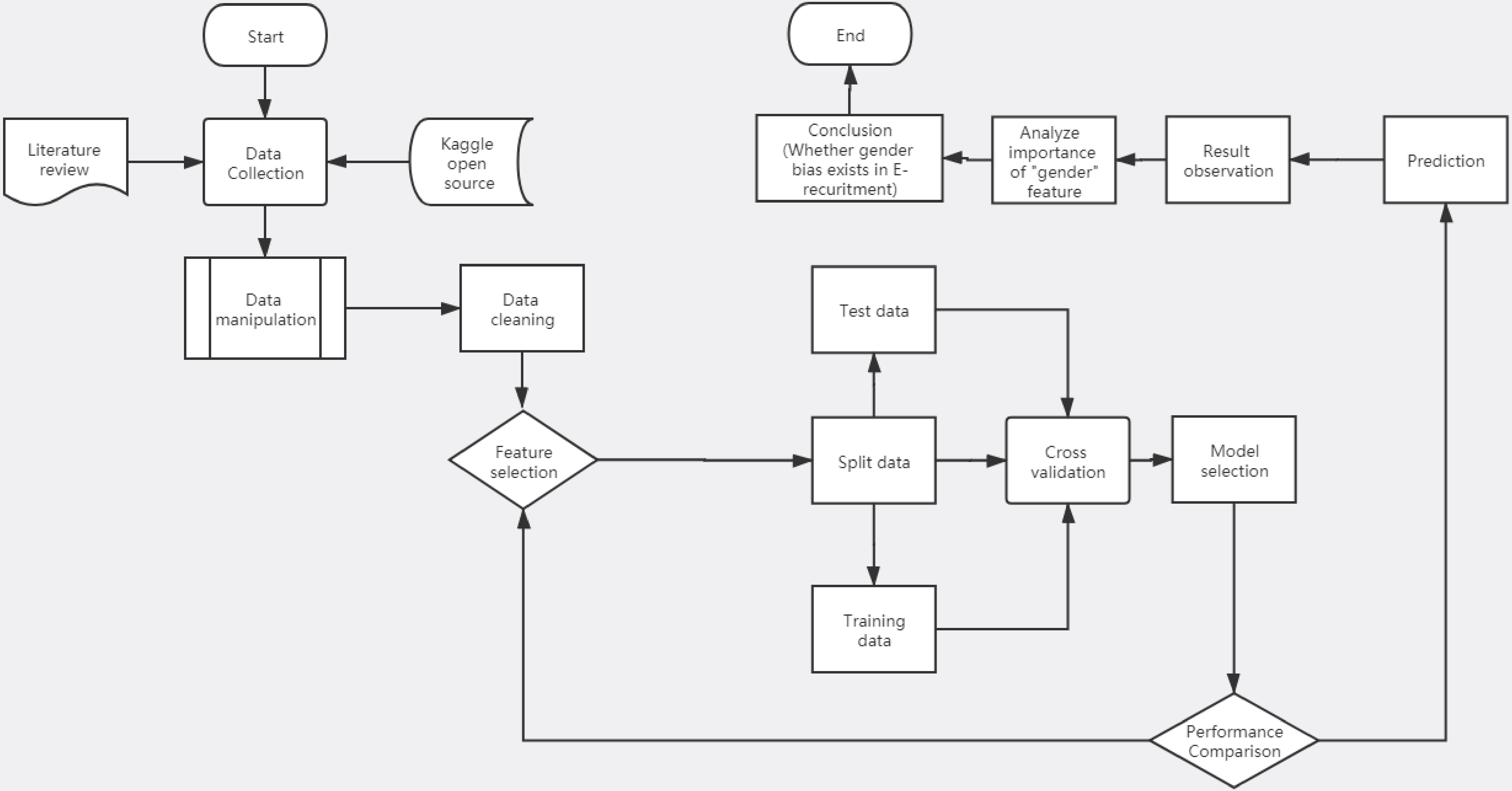
- focusing on features
- solve the problems of enterprises
- fulfill automatic match systems
- emphasize efficiency

previous computaional studies on E-recruitment tend to be utilitarian, besides help job-hunters to find a suitable job, there are more information we can study from data.

While traditional method of E-recruritment study highlight:

- social fairness
- opportunity equality
- bias in algorithm
- the situation of disadvantaged groups

# Frame work



Start
Literature review
Data Collection
Kaggle open source
Data manipulation
Data cleaning
Feature selection
Test data
Split data
Training data
Cross validation
Model selection
Performance Comparison
Prediction
Result observation
Analyze importance of "gender" feature
Conclusion (Whether gender bias exists in E-recuritment)
End

# Input-Data describe

Anonymous data from profiles scraped on LinkedIn. Contains data from about 60000 profiles with 52 features. Includes all their work history, gender, ethnicity, number of followers as well as analysis of their photo and name.

In original Data files each row contains:

- Profile data
- Job data (Company_name, Employment staus)
- Name analysis (Race, Gender, Age)
- Profile picture analysis (Attractiveness, Health, Emotionality)

However the dataset was unequal especially in ethnicity & gender:

| | | | |
|---|---|---|---|
| White | 76% | Asian | 16% |
| Male | 76% | Female | 24% |

# Data manipulation

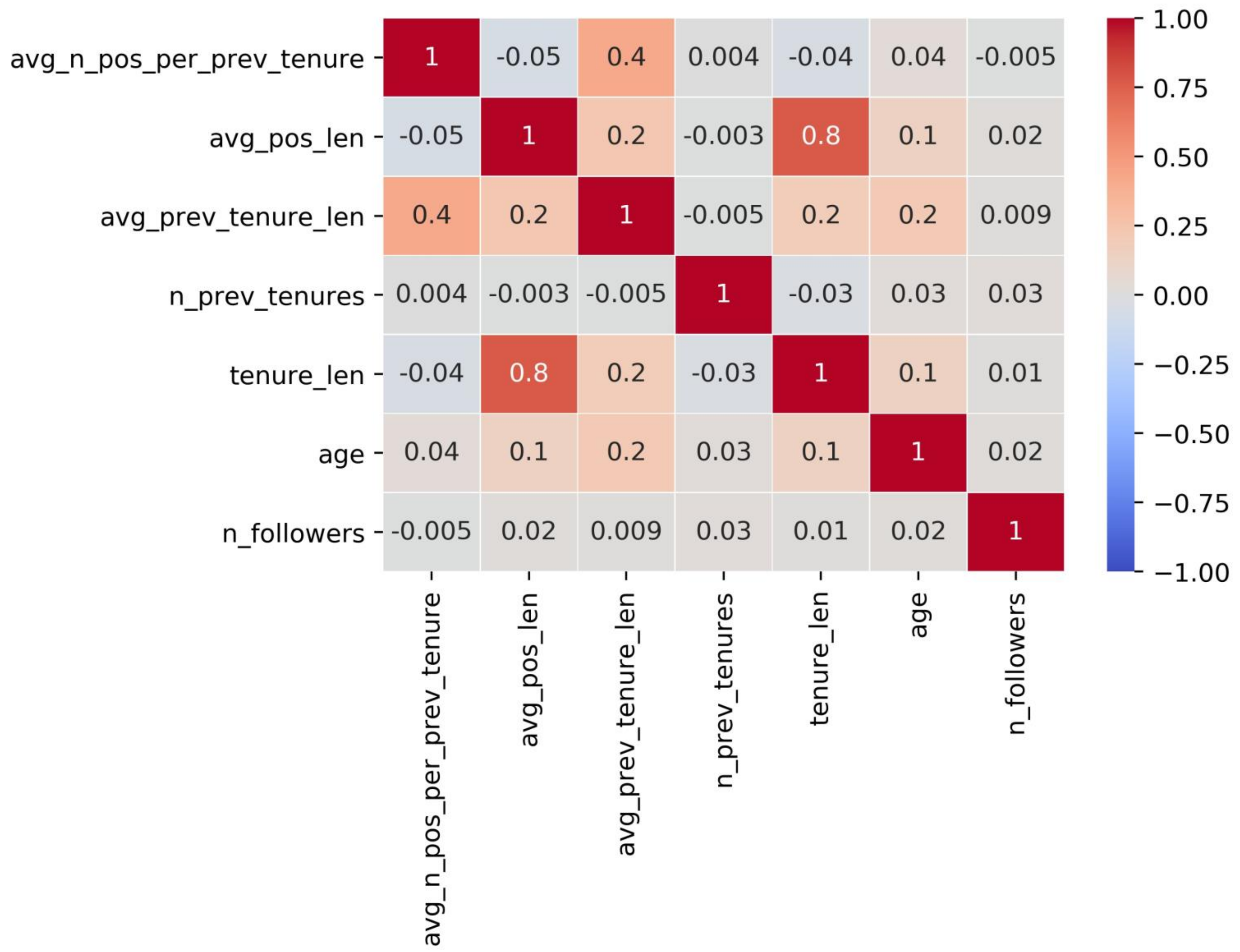
| | avg_n_pos_per_prev_tenure | avg_pos_len | avg_prev_tenure_len | n_prev_tenures | tenure_len | age | n_followers |
|---|---|---|---|---|---|---|---|
| **count** | 62674.000000 | 62674.000000 | 62674.000000 | 62674.000000 | 62674.000000 | 62674.000000 | 62612.000000 |
| **mean** | 1.194420 | 765.494312 | 1100.858561 | 4.302119 | 961.743434 | 44.071074 | 1226.329713 |
| **std** | 0.506835 | 750.559928 | 985.960617 | 38.679425 | 1088.256794 | 11.108183 | 6411.388029 |
| **min** | 1.000000 | -120.000000 | 0.000000 | 0.000000 | -120.000000 | 1.000000 | 0.000000 |
| **25%** | 1.000000 | 274.000000 | 537.000000 | 2.000000 | 304.000000 | 37.000000 | 405.000000 |
| **50%** | 1.000000 | 578.000000 | 882.800000 | 3.000000 | 640.000000 | 44.000000 | 725.000000 |
| **75%** | 1.200000 | 1035.000000 | 1400.000000 | 4.000000 | 1217.000000 | 51.000000 | 1244.000000 |
| **max** | 15.000000 | 21884.000000 | 39781.000000 | 4930.000000 | 21884.000000 | 791.000000 | 530566.000000 |

```
RangeIndex: 62674 entries, 0 to 62673
Data columns (total 11 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   avg_n_pos_per_prev_tenure  62674 non-null  float64
 1   avg_pos_len                62674 non-null  float64
 2   avg_prev_tenure_len        62674 non-null  float64
 3   n_pos                      62674 non-null  object
 4   n_prev_tenures             62674 non-null  int64
 5   tenure_len                 62674 non-null  int64
 6   age                        62674 non-null  float64
 7   ethnicity                  62674 non-null  object
 8   gender                     62674 non-null  object
 9   n_followers                62612 non-null  float64
 10  c_name                     62668 non-null  object
dtypes: float64(5), int64(2), object(4)
memory usage: 5.3+ MB
```

# Prior Observation
# -
# Why use nonlinear model

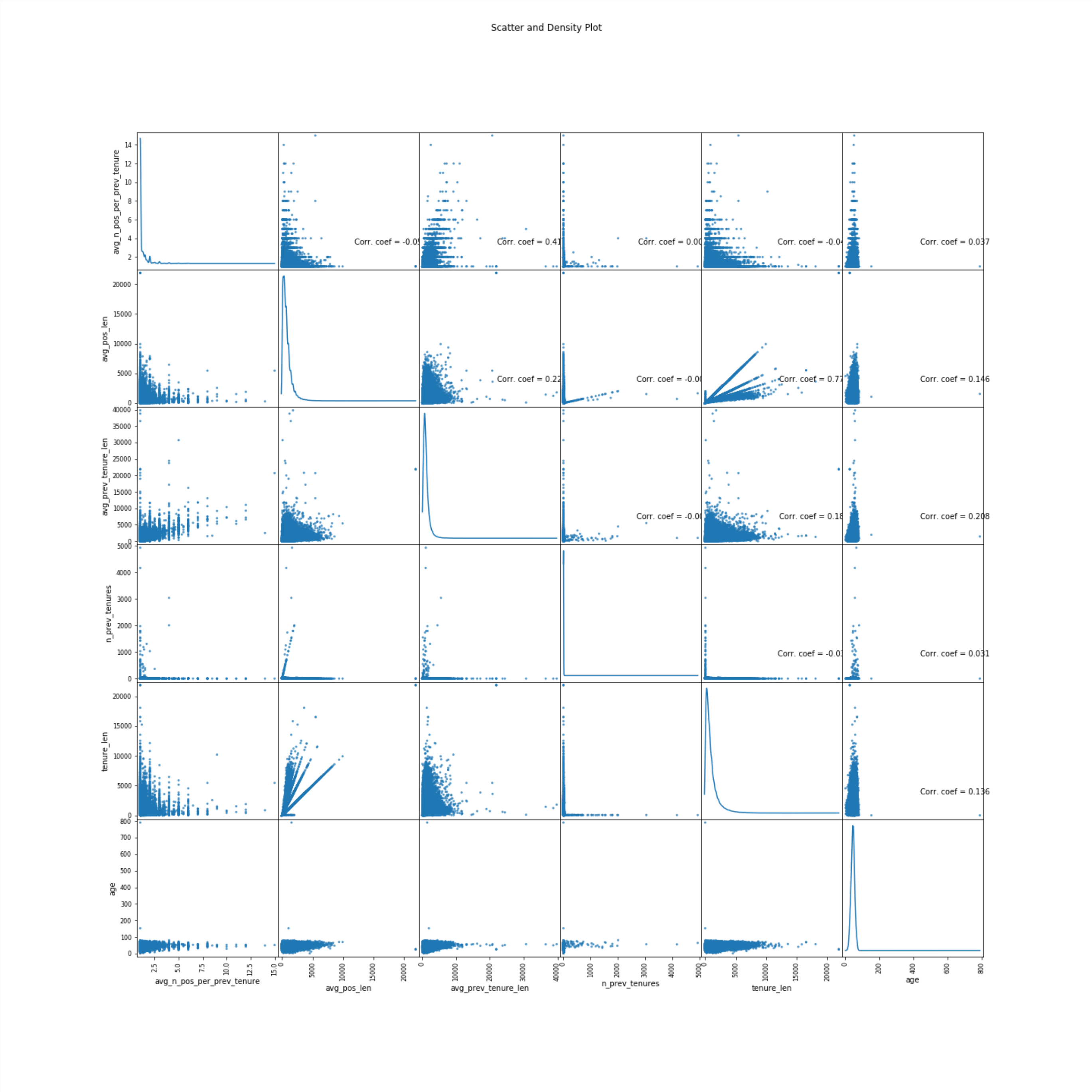
$$r_{xy} = \frac{Cov\ xy}{\sigma\ x\ \sigma y}$$

**Pearson's correlation**



# Prior Observation-Poisson Distribution

$$P(X=x)=e^{-u\frac{\mu^{x}}{x!}} \quad x\in R$$



# Future work-Data Wrangling and Cleaning

To observe whether gender inequality exists in top high-tech industry E-recruitment, we revised the data in 3 dimensions:

1) Reduce male profiles，especially those don't work in high-tech companies, until the proportion of male and female approaches the same.

2) Select target top high-tech corporations, which are Microsoft, IBM, Google, Adobe and some other similar companies.

3) One-hot encoding for object type features

- Binary Classification (0,1)

1) Try to use Decision Tree and Random Forest as model to predict the Probability of getting Job in Top high-tech Companies.

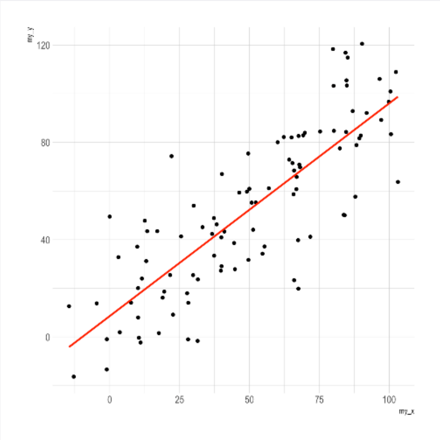
2) Analyze the significance of features and find out which variable affect model performance most, especially the gender bias.

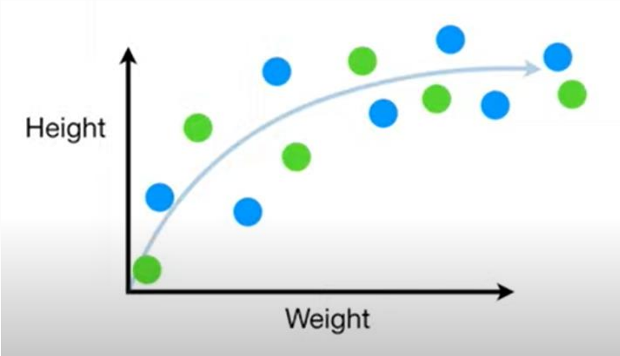
- Evaluation and validation

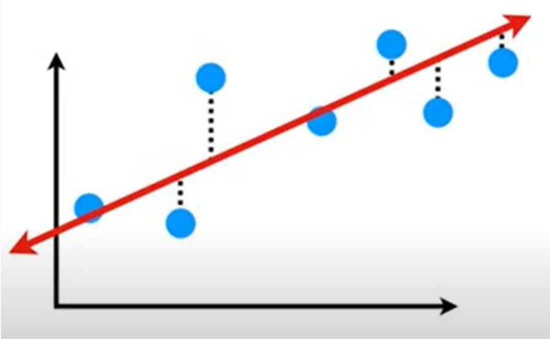
# Suggestion - Machine Learning Design and Algorithm (Recruiter’s perspective)

- **Design puzzle or riddle to test candidates aptitude and problems solving skills**
- **Interview candidate with machine learning algorithm (*HireVue, Inc.*)**
  - *assess their responses by using Natural Language Processing(NLP)*
  - *assess body language by machine vision*
- **Assess candidates’ soft skills using dictionary method**

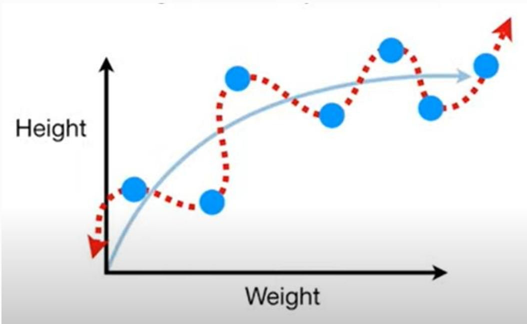
# Machine learning limitations



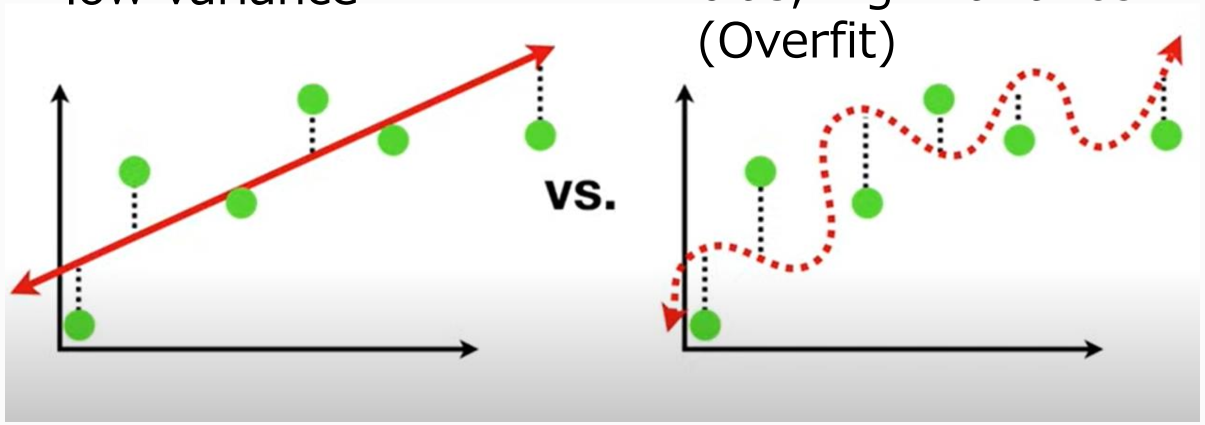


Linear: High bias, low variance




Wave line: Low bias, high variance (Overfit)




3 methods to find better fit: Regularization, Boosting and Bagging(random forest).

Photo source: *Wikipedia*

# Big data – limitation

- E-recruitment is 85% of all recruitment
- Data of what platform candidate obtained job offer is missing

# Thank you for listening!

Q&A