Predicting IBM Employee Attrition

Presented by R. Domadia, D. Padmanabhan, and U. Nigam

Data: IBM Developer

- Data published on Kaggle
- Roughly 1.4K observations including education, marital status, standard hours worked, monthly rate, and job-level
- Not a representative sample of all employee resignation from International Business Machines Corporation

Background

Employees Success = Organizational Success!!

- Backbone of an organization
- Expensive to train, both money and time
  - Turnover cost range from 30% to 150% of the employee's salary

Customers and Clients Trust = More Business...

- Departures take valuable knowledge to competitors
- Vacant position is expensive and results in a temporary loss in efficiency

Impact on Society and Economy

- Burnout costs the U.S. $190 billion in health care expenses
- 120,000 stress-attributed deaths
- Employee disengagement costs more than $500 billion per year

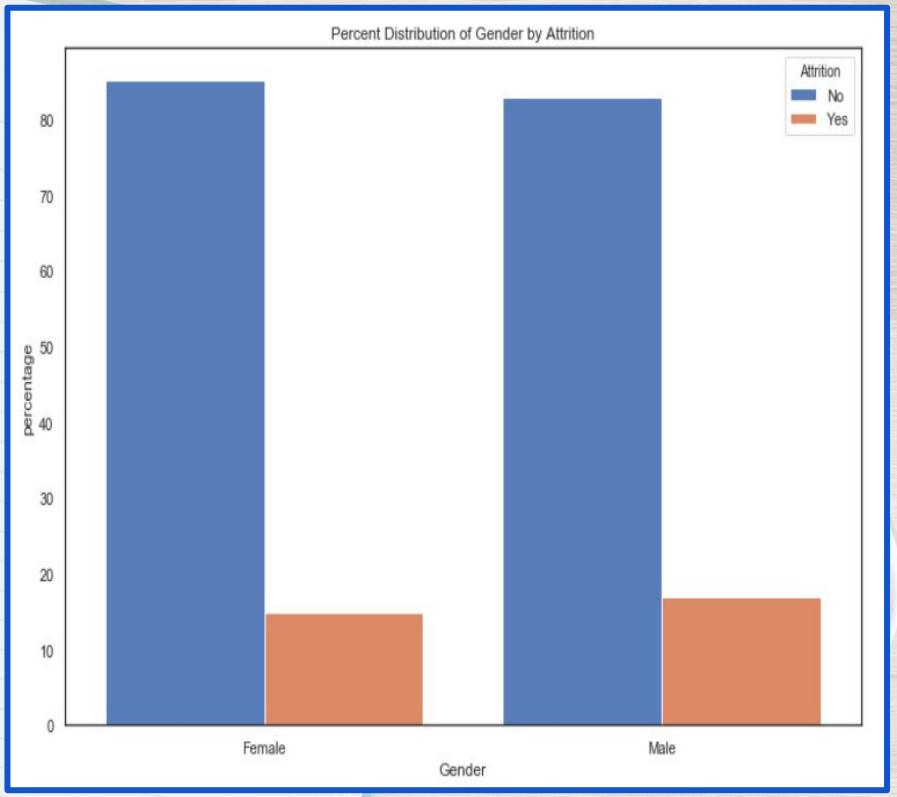
Glimpse of the Dataset

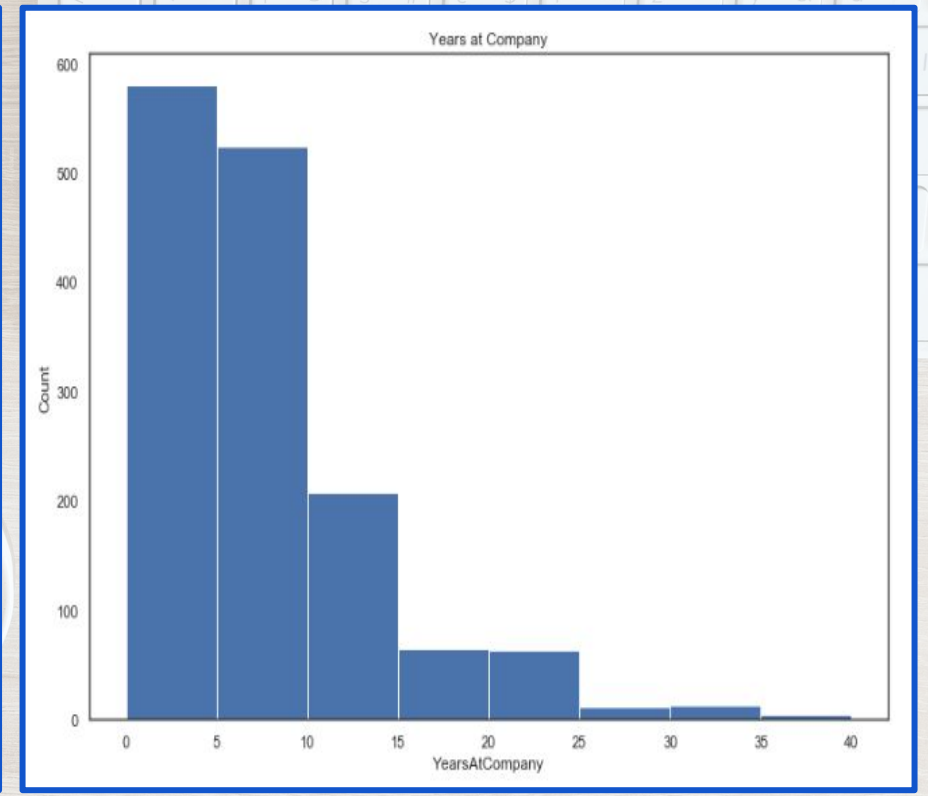
Attrition[Yes] means the employee is no longer employed at IBM

Attrition[No] means the employee is still employed with IBM

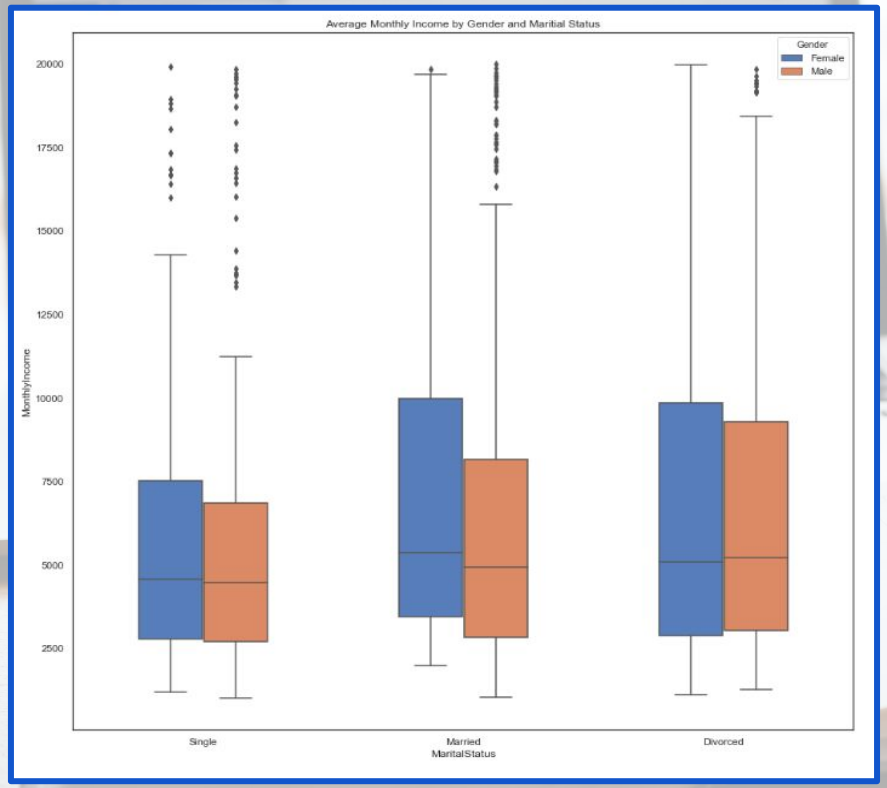
Age	Attrition	Department	DistanceFromHome	Education	Gender	JobSatisfaction	MaritalStatus
41	Yes	Sales	1	2	Female	4	Single
49	No	Research & Development	8	1	Male	2	Married
37	Yes	Research & Development	2	2	Male	3	Single
33	No	Research & Development	3	4	Female	3	Married
27	No	Research & Development	2	1	Male	2	Married
32	No	Research & Development	2	2	Male	4	Single
59	No	Research & Development	3	3	Female	1	Married
30	No	Research & Development	24	1	Male	3	Divorced
38	No	Research & Development	23	3	Male	3	Single
36	No	Research & Development	27	3	Male	3	Married

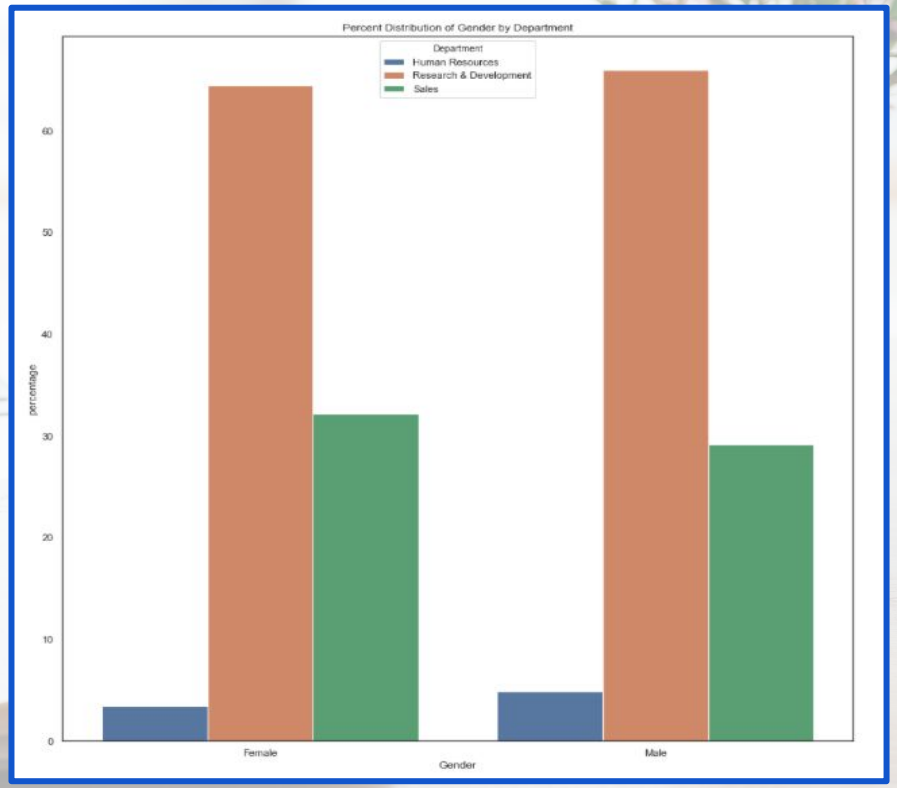
Descriptive Statistics: Attrition





Descriptive Statistics: Gender





SMART Question

What circumstances encourage IBM employees to resign?

Random Forest

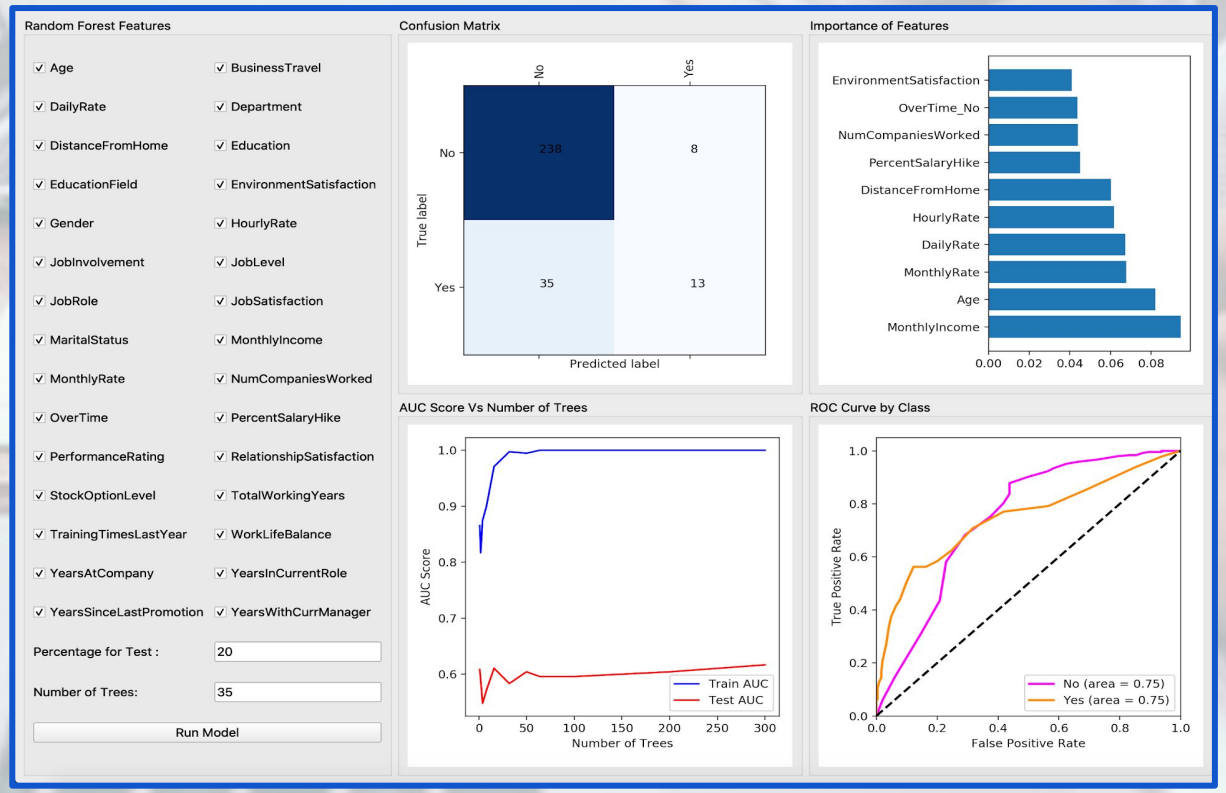
- Measures how much the accuracy decreases if the variable is excluded
- Measures the decrease of Gini impurity when a variable is chosen to split a node

Results:

	precision	recall	f1-score	support
0	0.87	0.97	0.92	246
1	0.62	0.27	0.38	48
accuracy			0.85	294
macro avg	0.75	0.62	0.65	294
weighted avg	0.83	0.85	0.83	294

Accuracy:

85.37414965986395



Decision Tree

- Comprehensive analysis on the outcome of a decision
- Missing values do not make a considerable impact

Results:

	precision	recall	f1-score	support
0	0.86	0.83	0.85	246
1	0.28	0.33	0.30	48
accuracy			0.75	294
macro avg	0.57	0.58	0.58	294
weighted avg	0.77	0.75	0.76	294

Accuracy:

75.17006802721087

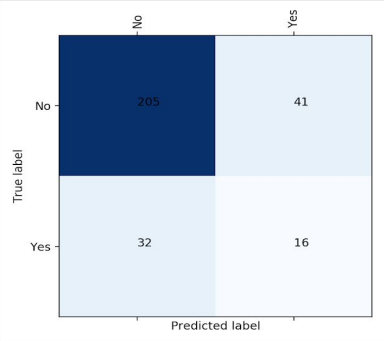
Decision Tree Features

- ☑ Age
- ☑ BusinessTravel
- ☑ DailyRate
- ☑ Department
- ☑ DistanceFromHome
- ☑ Education
- ☑ EducationField
- ☑ EnvironmentSatisfaction
- ☑ Gender
- ☑ HourlyRate
- ☑ JobInvolvement
- ☑ JobLevel
- ☑ JobRole
- ☑ JobSatisfaction
- ☑ MaritalStatus
- ☑ MonthlyIncome
- ☑ MonthlyRate
- ☑ NumCompaniesWorked
- ☑ OverTime
- ☑ PercentSalaryHike
- ☑ PerformanceRating
- ☑ RelationshipSatisfaction
- ☑ StockOptionLevel
- ☑ TotalWorkingYears
- ☑ TrainingTimesLastYear
- ☑ WorkLifeBalance
- ☑ YearsAtCompany
- ☑ YearsInCurrentRole
- ☑ YearsSinceLastPromotion
- ☑ YearsWithCurrManager

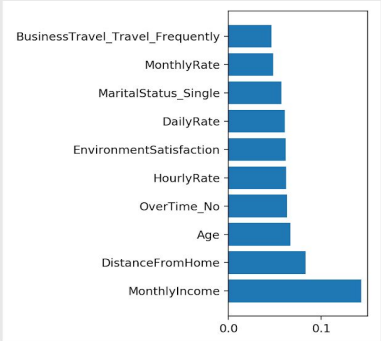
Percentage for Test : 20

Run Model

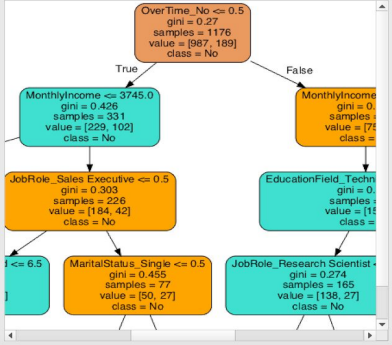
Confusion Matrix



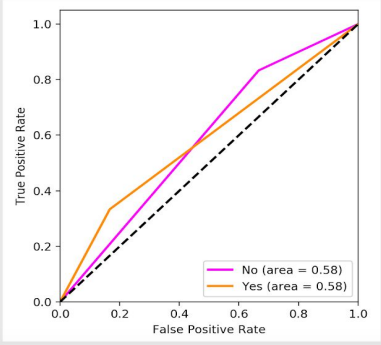
Importance of Features



Decision Tree Graph



ROC Curve by Class



Logistic Regression

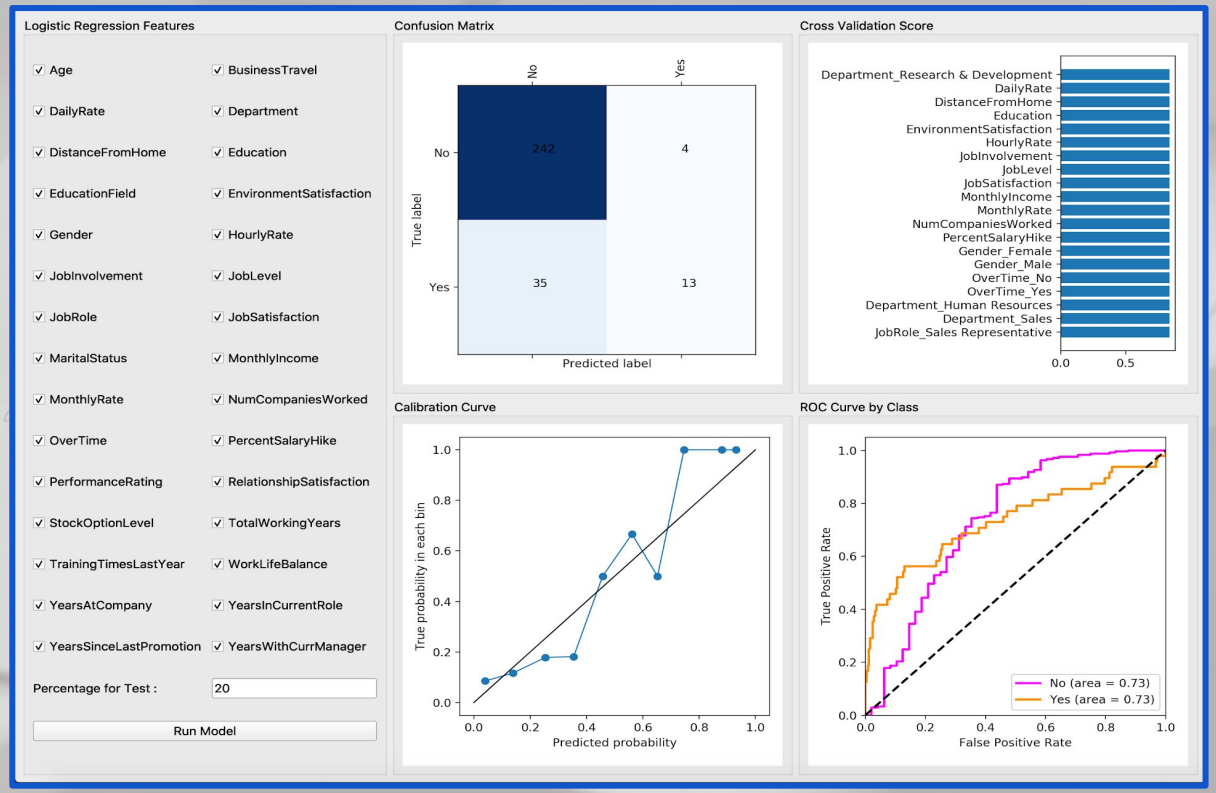
- Explains the relationship between two variables

Results:

	precision	recall	f1-score	support
0	0.87	0.98	0.93	246
1	0.76	0.27	0.40	48
accuracy			0.87	294
macro avg	0.82	0.63	0.66	294
weighted avg	0.86	0.87	0.84	294

Accuracy:

86.73469387755102



K-Nearest Neighbors

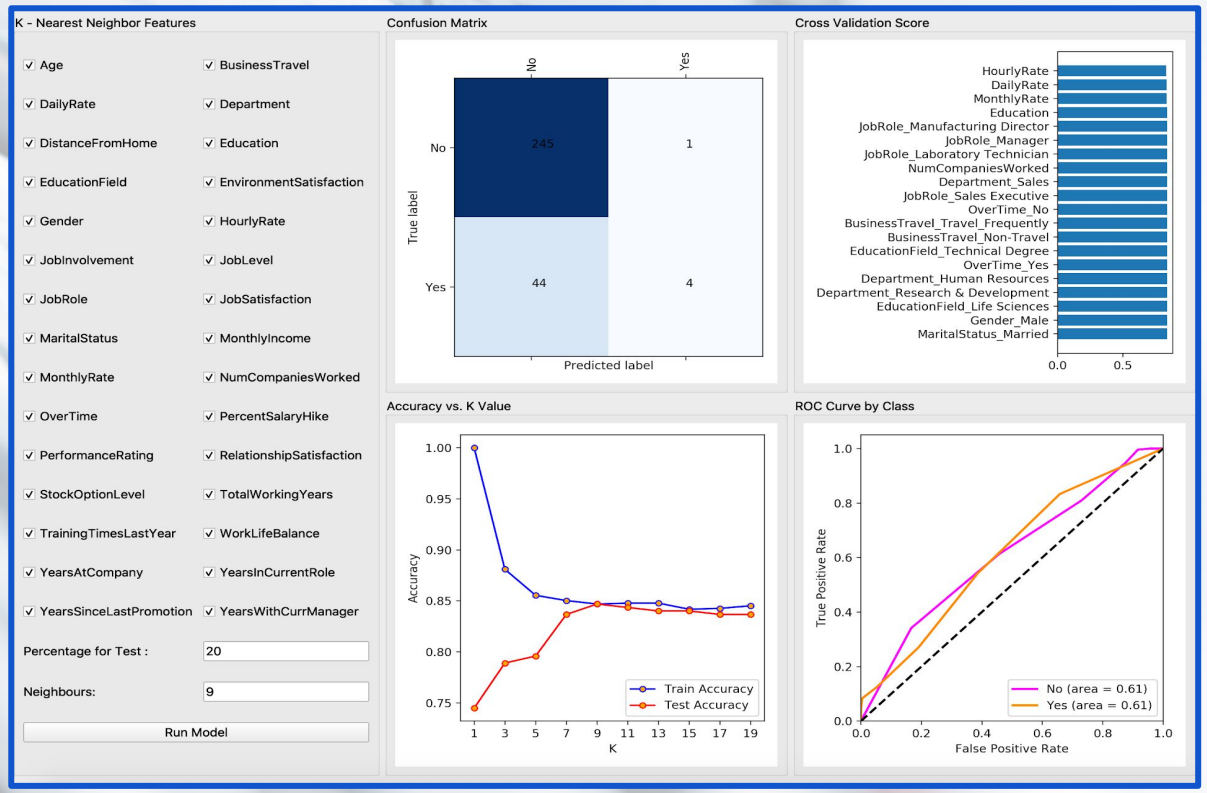
- Assumes that similar things exist in close proximity.
- Supervised Clustering algorithm

Results:

	precision	recall	f1-score	support
0	0.85	1.00	0.92	246
1	0.80	0.08	0.15	48
accuracy			0.85	294
macro avg	0.82	0.54	0.53	294
weighted avg	0.84	0.85	0.79	294

Accuracy:

84.6938775510204



Fun Facts

57% of organizations view employee retention as a problem, and 22% of new hires leave their jobs within 45 days of being hired.