Job Market Consulting

2024

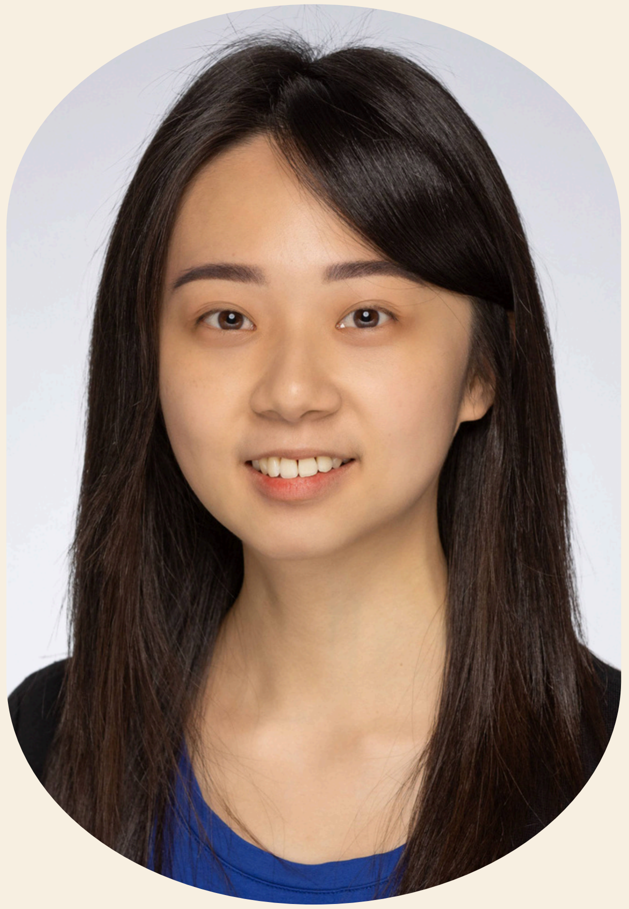
Talent Optima

By Félix Veaux, Lincoln, Shuxi, Maralma

Our Team

Felix Veaux

FelixVeaux



Shuxi Chen

shuxi-ch

Maral Batnasan

Maral-Batnasan

Xuanlong(Lincoln) Lyu

Lincolnlyu

LinkedInJobPosting

github.com/McGill-MMA-EnterpriseAnalytics/LinkedInJobPosting

Table of Contents

TalentOptima HR Consulting

Optimizing Job Postings & Maximizing Career Value with Data-Driven Insights

The Problem

For Companies

- Companies have tasks that need to be done and need labour to do them.
- Nearly 70% of job seekers consider salary transparency critical (Glassdoor, 2022) yet only ~35% of job postings include salary data. source
- Average time-to-hire is around 42 days (SHRM), indicating inefficiencies in traditional hiring processes. source
- 87.6% of Human Resources (HR) managers report using salary benchmarks to set pay. source

For Job Seekers

- There is a Lack of transparent market data, leaving candidates vulnerable to undervalued compensation offers.
- Over 65% of candidates report difficulty in benchmarking salaries without reliable data. source

Our Solution

We offer comprehensive solutions for companies and individuals. Our solutions are at the core of our service and bring data-driven insights to customers.

ROBUST ML INSIGHTS

NLP TO EXTRACT NUANCED INSIGHT

DATA-DRIVEN SALARY RANGES

DATA INSIGHTS TO REDUCE TIME-TO-HIRE

Business Model & Competitive Advantage

Software as a Service (SaaS)

1

2

3

Revenue Streams

Consulting services, subscription-based SaaS platform, and premium analytics packages.

High Margins

SaaS models in HR tech typically enjoy recurring revenues and gross margins exceeding 70%. source

Competitive Edge

Our dual approach, targeting employers and job seekers and leveraging advanced ML (NLP, ensemble methods), differentiates us from traditional HR consultancies.

Market Oportunity

TAM
(Total Addressable Market)

$16.4 billion
global HR technology market

SAM
(Serviceable Addressable Market)

$0.5 billion
recruitment analytics and job posting optimization

SOM
(Serviceable Obtainable Market)

$25 Million
5% SAM within 5 years

source

Traction & Roadmap

Q1
DATASET MERGE

Successfully merged and processed data from 44,000+ job postings—More scraping will be done to improve this.

Q2
MODEL PROTOTYPE

Early model prototypes show promising improvements in predictive accuracy and engagement forecasting

Q3
PILOT TESTING

Pilot testing with key clients, model refinement, and beta launch within the next 6 months.

Q4
ITERATION

Enhance our model through continuous improvement and pilot program feedback.

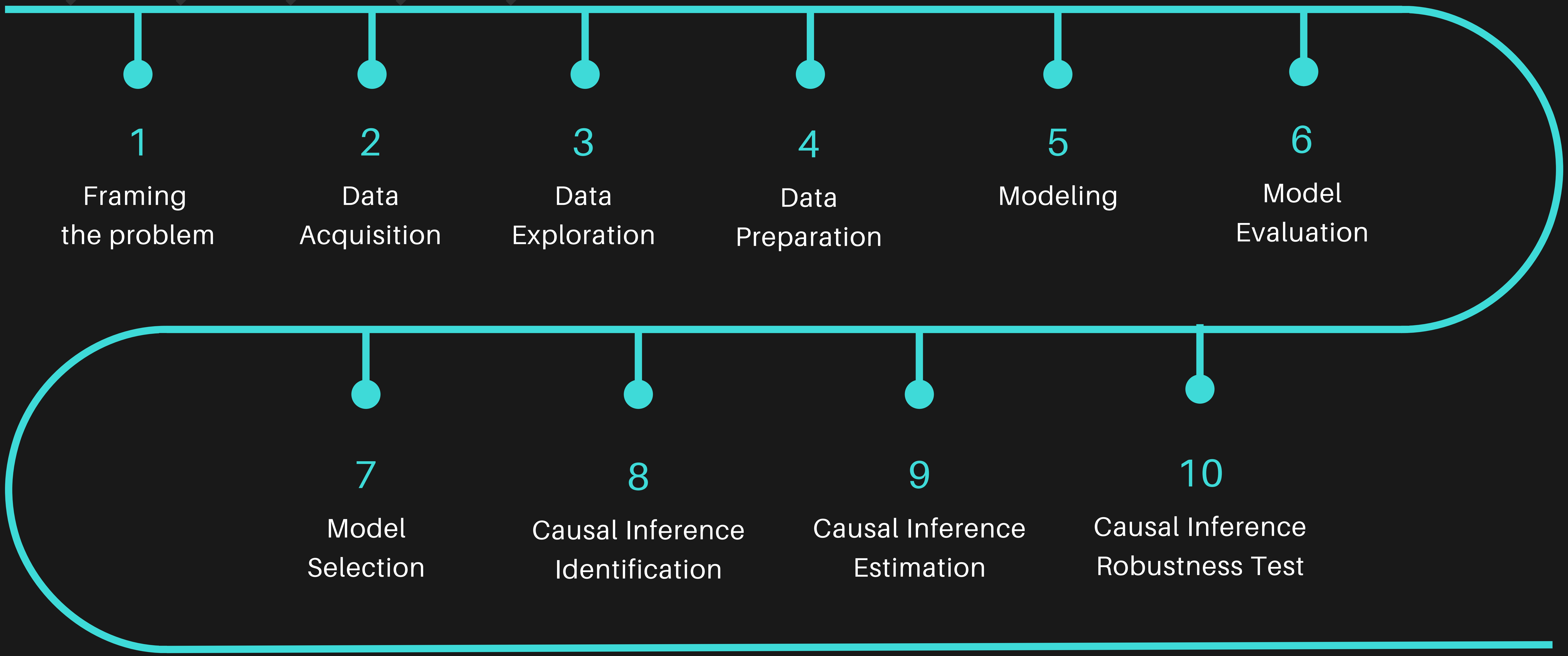
Q5
PRODUCT LAUNCH

Launch the product with a strategic marketing campaign to boost visibility and attract customers.

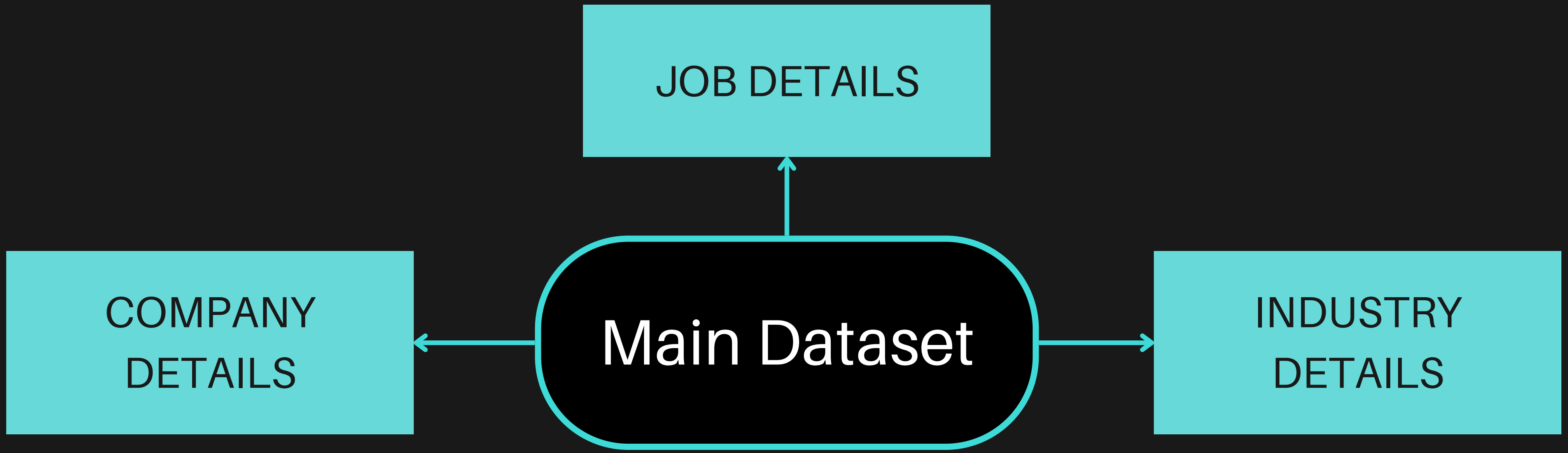
Q6
GROWTH AND SCALING

Scaling the platform, expanding market segments, and forging strategic partnerships to capture a larger share of the HR tech market.

Data Science Lifecycle



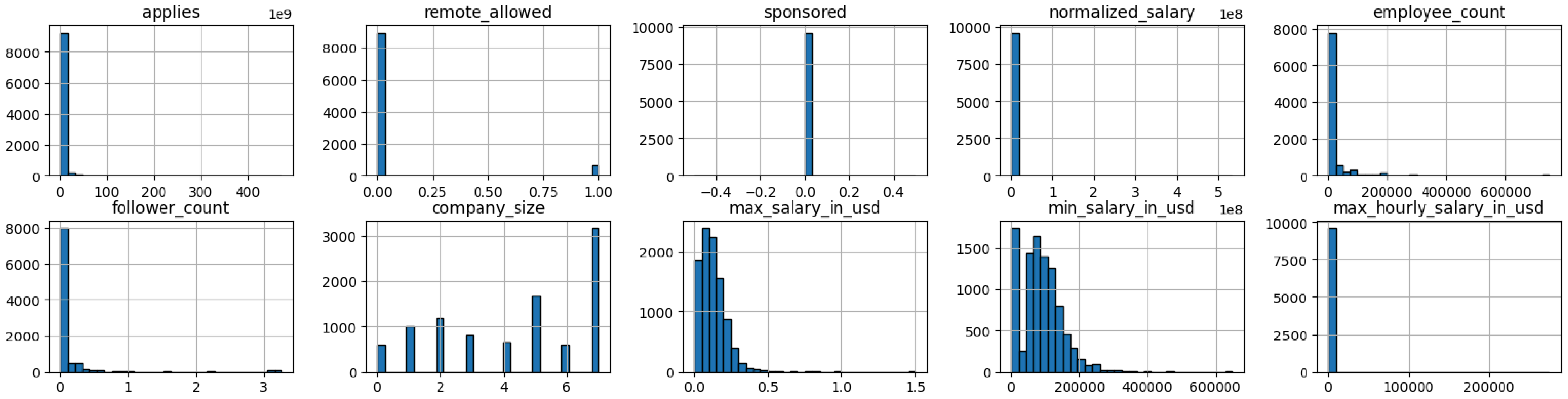
2 DATA ACQUISITION



NUMERICAL PREDICTORS	CATEGORICAL PREDICTORS	BINARY PREDICTORS

• Maximum salary
• Minimum salary
• Normalized salary
 	
• Number of times the job posting has been viewed
• Number of applications that have been submitted
 	
• Number of employees at company
• Number of company followers on LinkedIn
	• Job experience level
	  ◦ (entry, associate, executive)
	• Job Title
	  ◦ (Extracted from Job Description using TF-IDF)
	• Job Location
	• Application Type (onsite, offsite)
		• Remote allowed
		• Child care support
		• Commuter benefits
		• Dental insurance
		• Disability insurance
		• Medical insurance
		• Paid maternity leave
		• Paid paternity leave
		• Pension plan
		• Student loan assistance
		• Tuition assistance
		• Vision insurance

3 DATA EXPLORATION



4

DATA PREPARATION

NEW FEATURES CREATED

- **Salary difference percentage** (between Min and Max salary)
- **Views and Applies per day**

NUMERICAL PREDICTORS	CATEGORICAL PREDICTORS
• Log transformation, outlier detection applied on employee count, follower count, normalized salary (for classification model)	• Job Title (NLP and TF-IDF) • Skills (NLP and TF-IDF) • State/City (One Hot Encode)

SMOTE

SMOTEEN

SMOTETETOMEK

5

MODELING

A. Classification Model - Engagement Level Prediction

Random Forest Classification

SVM

6

MODEL EVALUATION

Metric	High Engagement	Low Engagement
Precision	0.71	0.74
Recall	0.60	0.82
F1-score	0.65	0.78

RF ACCURACY: 0.73

Metric	High Engagement	Low Engagement
Precision	0.58	0.71
Recall	0.65	0.65
F1-score	0.61	0.68

SVM ACCURACY: 0.65

7 MODEL SELECTION

Random Forest Classification

8 MODEL FINE-TUNING

RANDOM FOREST – GRID SEARCH

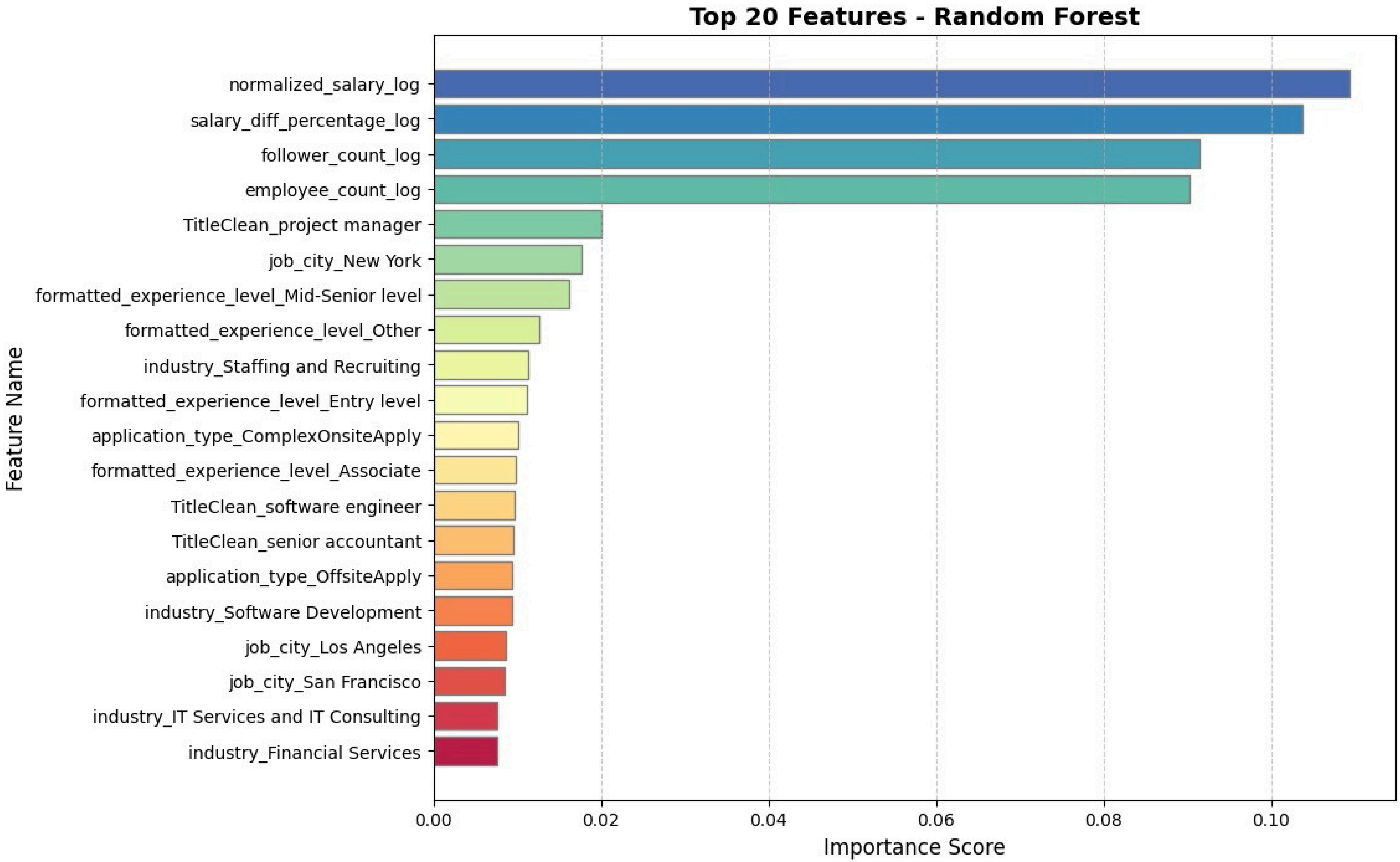
Metric	High Engagement	Low Engagement
Precision	0.76	0.76
Recall	0.64	0.85
F1-score	0.70	0.81

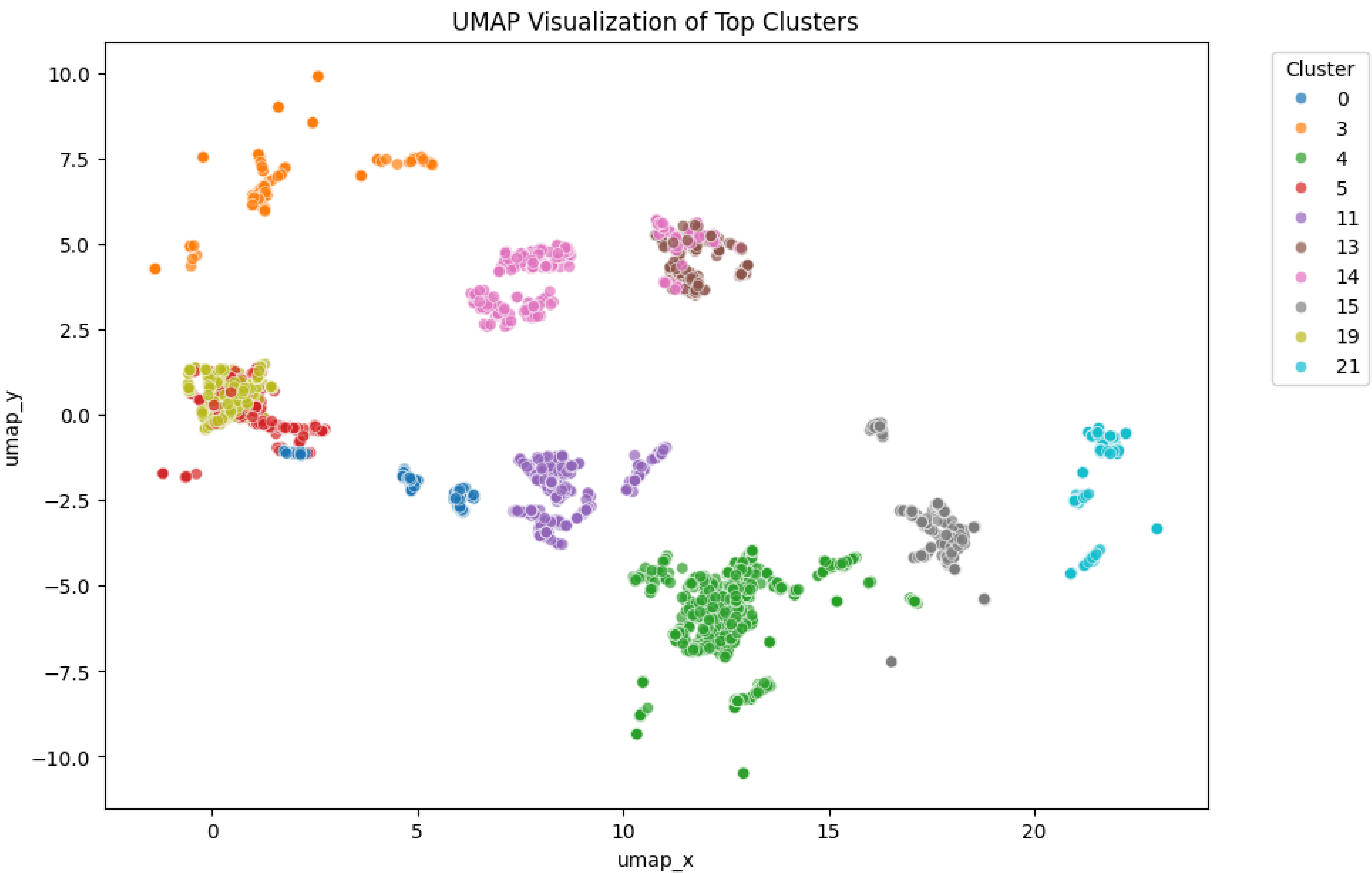
RF ACCURACY: 0.76

VOTING CLASSIFIER: RANDOM FOREST AND SVM

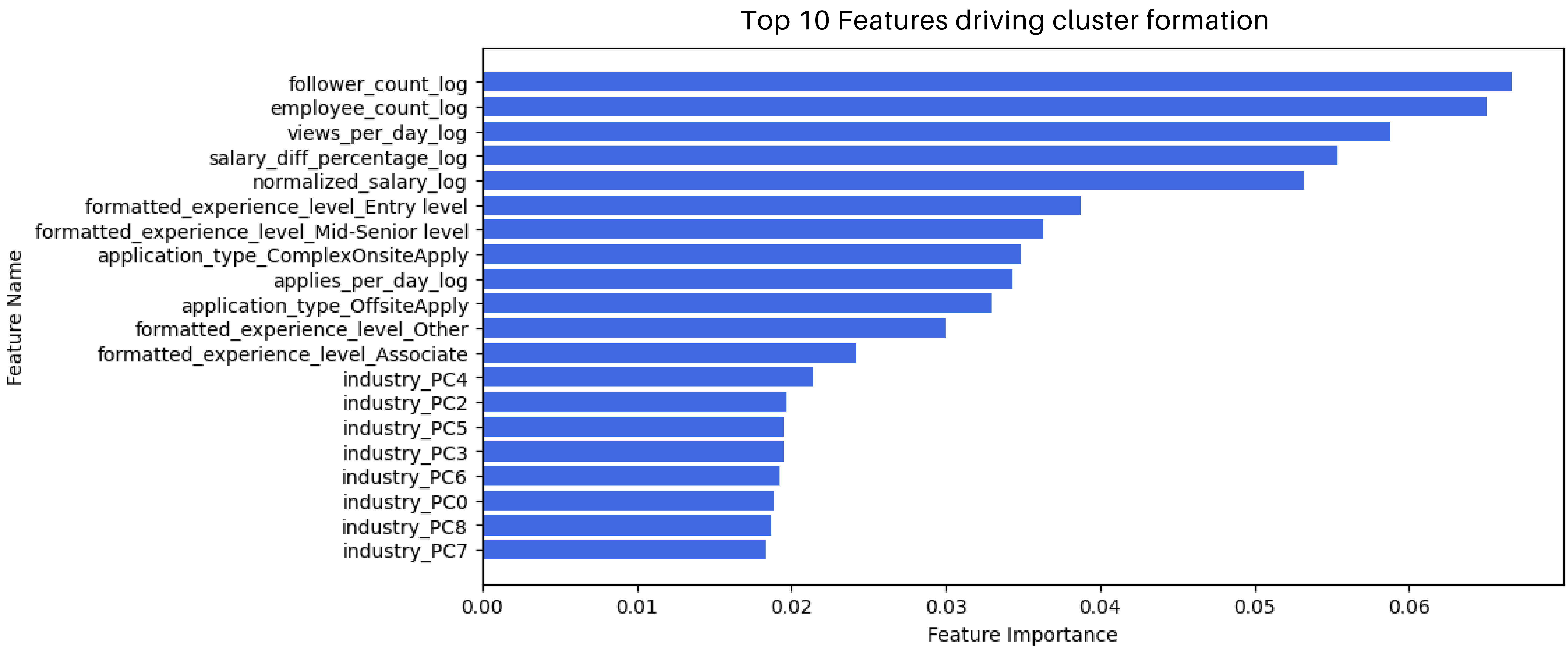
Metric	High Engagement	Low Engagement
Precision	0.60	0.78
Recall	0.76	0.62
F1-score	0.67	0.69

RF ACCURACY: 0.68


Top 20 Features - Random Forest
normalized_salary_log
salary_diff_percentage_log
follower_count_log
employee_count_log
TitleClean_project manager
job_city_New York
formatted_experience_level_Mid-Senior level
formatted_experience_level_Other
industry_Staffing and Recruiting
formatted_experience_level_Entry level
application_type_ComplexOnsiteApply
formatted_experience_level_Associate
TitleClean_software engineer
TitleClean_senior accountant
application_type_OffsiteApply
industry_Software Development
job_city_Los Angeles
job_city_San Francisco
industry_IT Services and IT Consulting
industry_Financial Services
Feature Name
0.00
0.02
0.04
0.06
0.08
0.10
Importance Score


UMAP Visualization of Top Clusters
umap_y
umap_x
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
0
5
10
15
20
Cluster
0
3
4
5
11
13
14
15
19
21

EXPLORING CLUSTER FORMATION



5

MODELING

Drop High Correlated Features

↓

Feature Selection

↓

Fine Tuning

6

MODEL SELECTION

Deep Feed-Forward Neural Network

Decision Tree

Gradient Boosting

7

MODEL EVALUATION

	Model	Test RMSE	Test R Square
Max Salary	Deep Feed-Forward Neural Network	72394.874515	0.487176
	Decision Tree	77848.260640	0.407006
	Gradient Boosting	0.507990	0.103441

7

MODEL EVALUATION

	Model	Test RMSE	Test R Square
Min Salary	Deep Feed-Forward Neural Network	45535.034424	0.500267
	Decision Tree	46888.618490	0.470115
	Gradient Boosting	59888.620254	0.135559

7

MODEL EVALUATION

	Model	Test RMSE	Test R Square
Hourly Salary	Deep Feed-Forward Neural Network	1551.718080	0.111550
	Decision Tree	2320.786726	0.013887
	Gradient Boosting	1710.365083	0.079407

8

CAUSAL INFERENCE

Identification

confounders

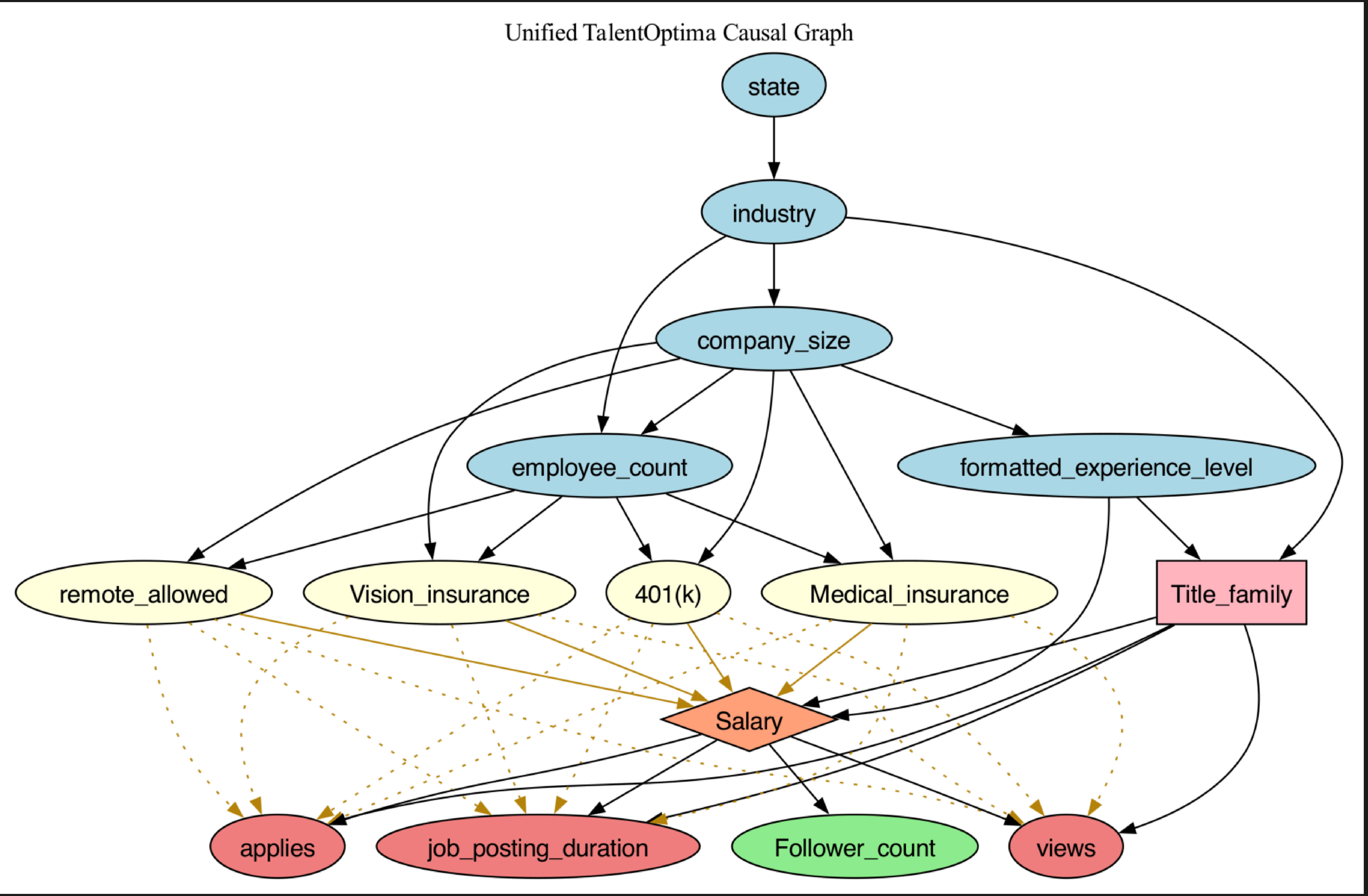
Treatments (Intervention)

Treatments (Intervention)

mediator

Post-treatment factors (backdoor > 1)

Post-treatment factors (backdoor = 1)



CAUSAL INFERENCE

8

Identification

confounders

Treatments (Intervention)

Treatments (Intervention)

mediator

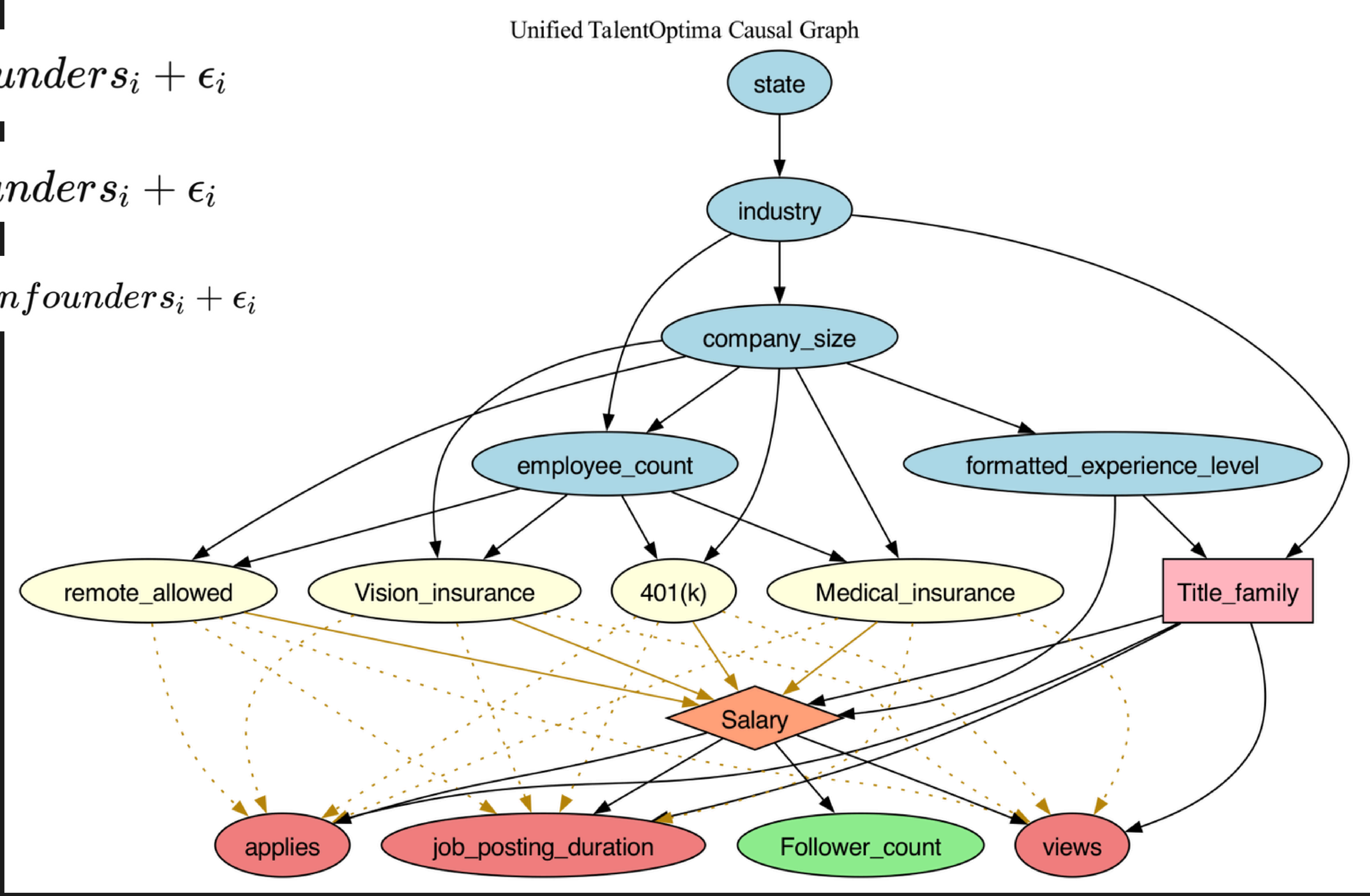
Post-treatment factors (backdoor = 1)

Post-treatment factors (backdoor > 1)

$$Applies_i = \alpha_0 + \alpha_1 \cdot Benefit_i + \alpha_2 \cdot Confounders_i + \epsilon_i$$

$$Salary_i = \beta_0 + \beta_1 \cdot Benefit_i + \beta_2 \cdot Confounders_i + \epsilon_i$$

$$Applies_i = \gamma_0 + \gamma_1 \cdot Salary_i + \gamma_2 \cdot Benefit_i + \gamma_3 \cdot Confounders_i + \epsilon_i$$



Total Effect of Benefits (α_1)

= (benefits -->Salary --> applies) + (benefits --> applies)

= **(Solid brown arrows) + (dotted brown arrows)**

= **(Indirect effect) + (Direct Effect)**

= **(β_1*gamma1) + gamma2**

9 CAUSAL INFERENCE

Estimation

Standardized Mean Difference (SMD)

&

Propensity Score Matching (PSM)

```
==================================================
Analyzing benefit: remote_allowed
==================================================

SMD Before Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 23
Large imbalance (SMD > 0.25): 2

SMD After Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 0
Large imbalance (SMD > 0.25): 0

Matched Sample: 3995 treated, 3995 control
```

```
==================================================
Analyzing benefit: 401(k)
==================================================

SMD Before Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 7
Large imbalance (SMD > 0.25): 0

SMD After Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 0
Large imbalance (SMD > 0.25): 0

Matched Sample: 7577 treated, 7577 control
```

```
==================================================
Analyzing benefit: Medical insurance
==================================================

SMD Before Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 15
Large imbalance (SMD > 0.25): 4

SMD After Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 0
Large imbalance (SMD > 0.25): 0

Matched Sample: 4111 treated, 4111 control
```

```
==================================================
Analyzing benefit: Vision insurance
==================================================

SMD Before Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 16
Large imbalance (SMD > 0.25): 2

SMD After Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 0
Large imbalance (SMD > 0.25): 0

Matched Sample: 4014 treated, 4014 control
```

9

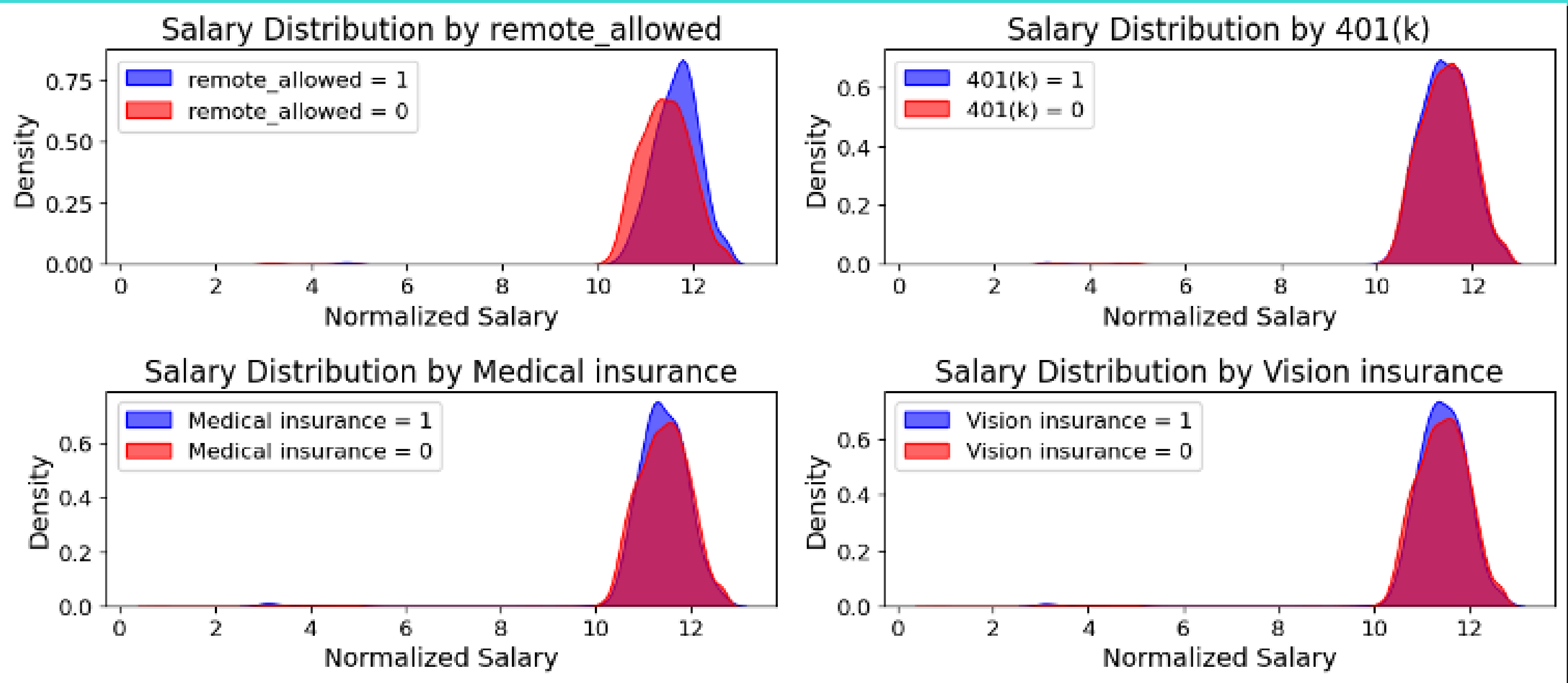
CAUSAL INFERENCE

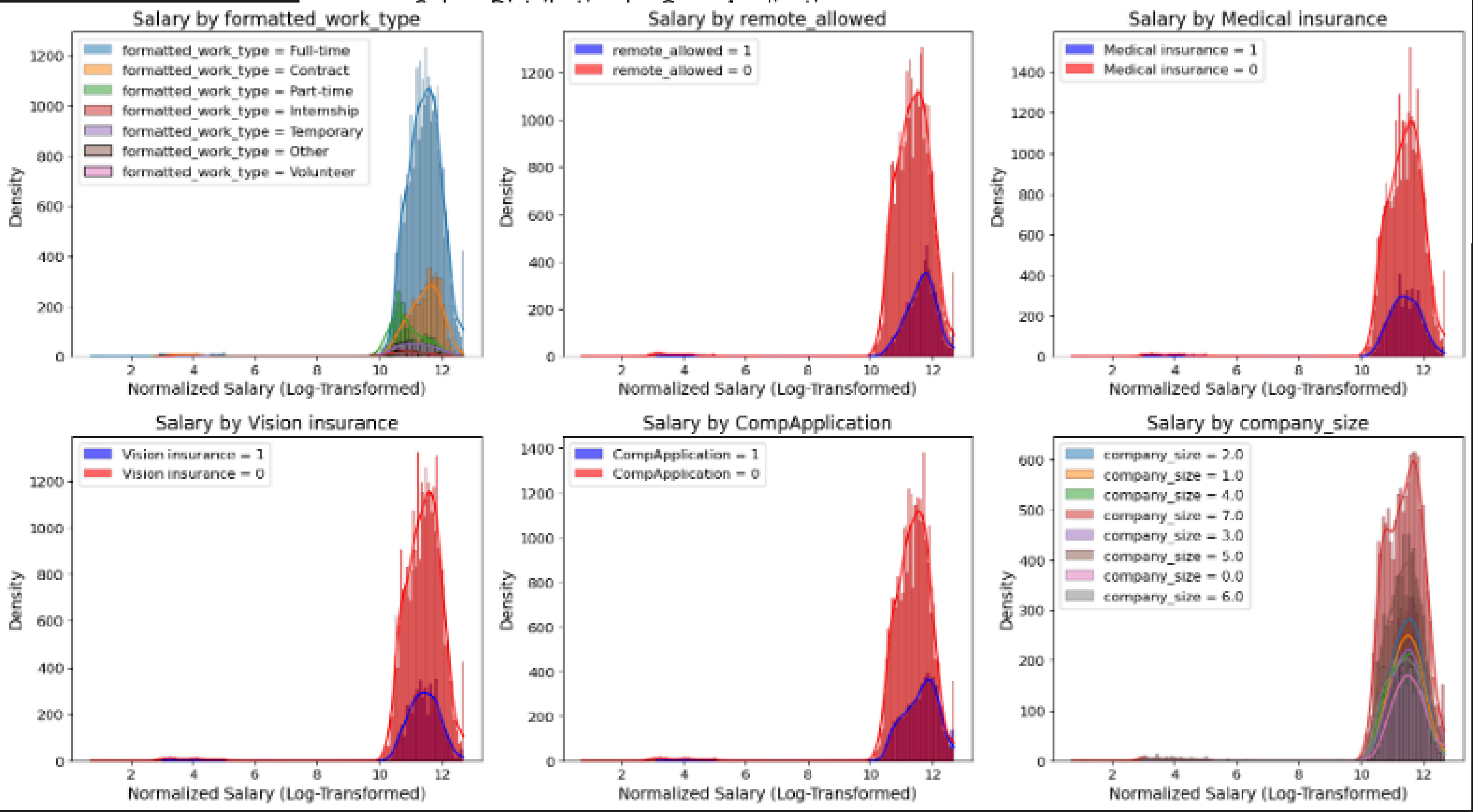
Estimation

Standardized Mean Difference (SMD)

&

Propensity Score Matching (PSM)





CAUSAL INFERENCE

9

Estimation

OLS - ATE

Summary of All Benefits Analysis:

Benefit	Direct Effect	Indirect Effect	Total Effect
remote_allowed	0.5776249367842361	-0.0023801389926877176	0.5752447977915565
401(k)	0.01811189355879506	0.0031359116185031835	0.021247805177299572
Medical insurance	0.042192981626901056	0.0065570627642114665	0.04875004439111449
Vision insurance	0.055923020913214085	0.00016062615305846738	0.056083647066268616

Remote >

Vision insurance >

Medical insurance >

remote_allowed

9

CAUSAL INFERENCE

Estimation

OLS - ATE

TLearner + XGBoost - CATE

Summary of All Benefits Analysis:

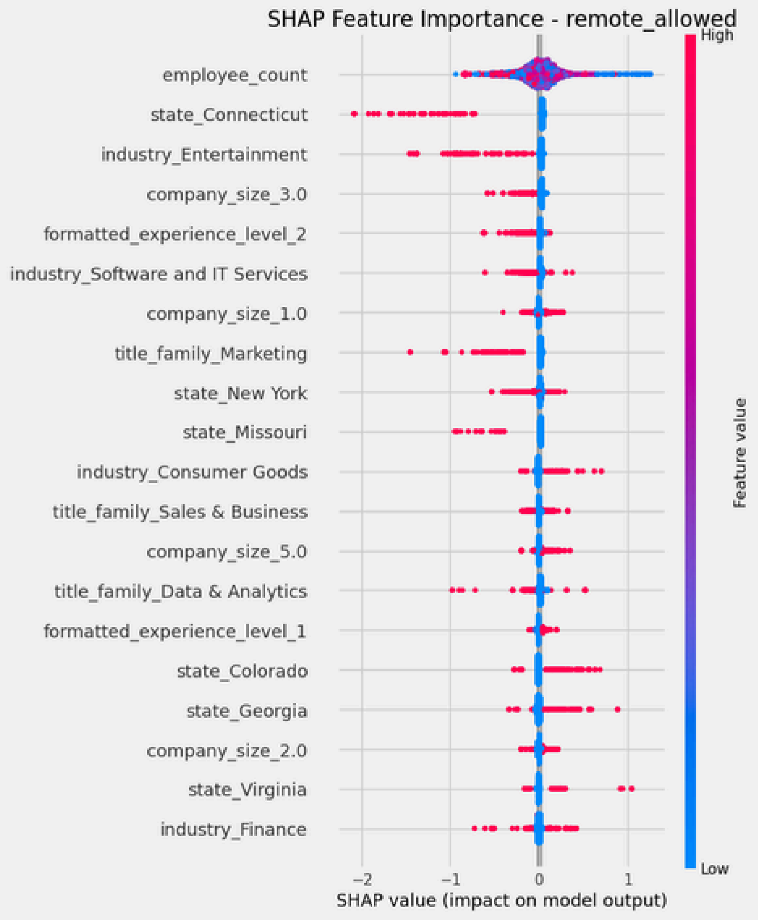
Benefit	Direct Effect	Indirect Effect	Total Effect	Mean CATE	Sample Size (T/C)	Imbalanced Covariates (Before→After)
remote_allowed	0.5776249367842361	-0.0023801389926877176	0.5752447977915565	0.5975475767900111	3995 / 3995	23 → 0
401(k)	0.01811189355879506	0.0031359116185031835	0.021247805177299572	0.02775046644542822	7577 / 7577	7 → 0
Medical insurance	0.04219298162690105 6	0.0065570627642114665	0.04875004439111449	0.06406321586353232	4111 / 4111	15 → 0
Vision insurance	0.05592302091321408 5	0.00016062615305846738	0.05608364706626861 6	0.061334590274938275	4014 / 4014	16 → 0

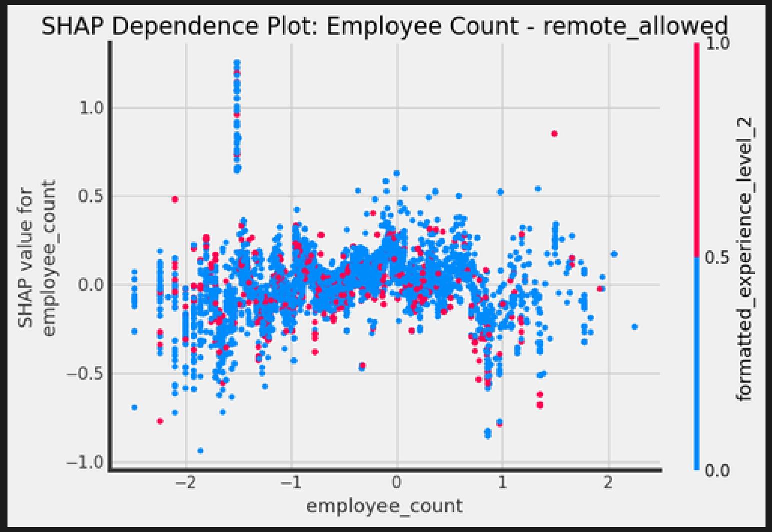
Remote > Vision insurance > Medical insurance > remote_allowed

9

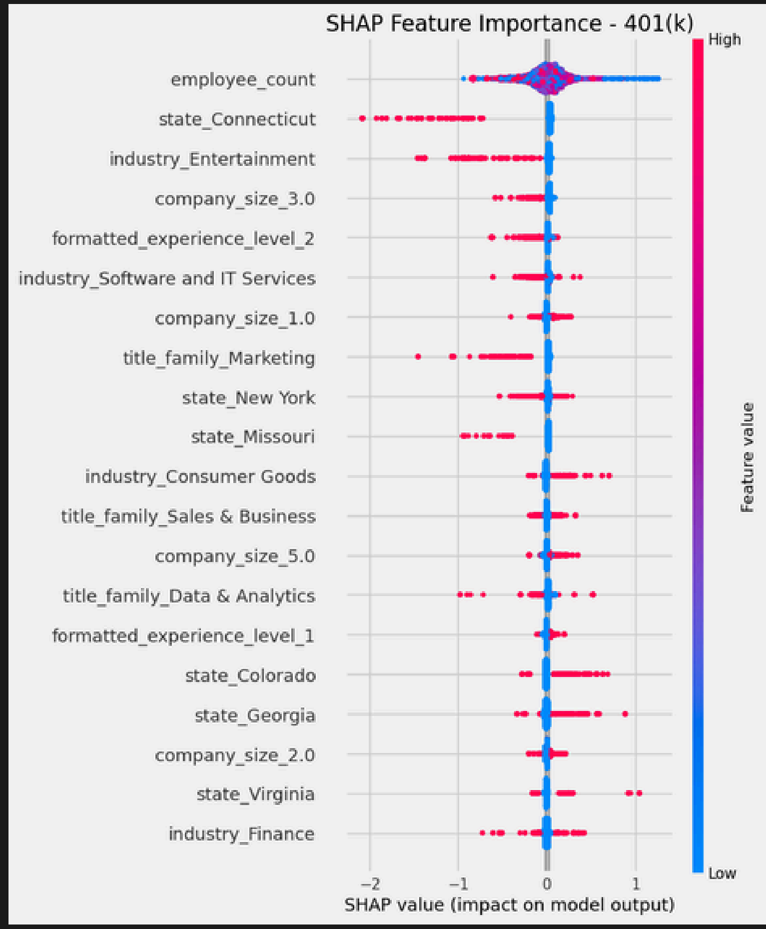
CAUSAL INFERENCE

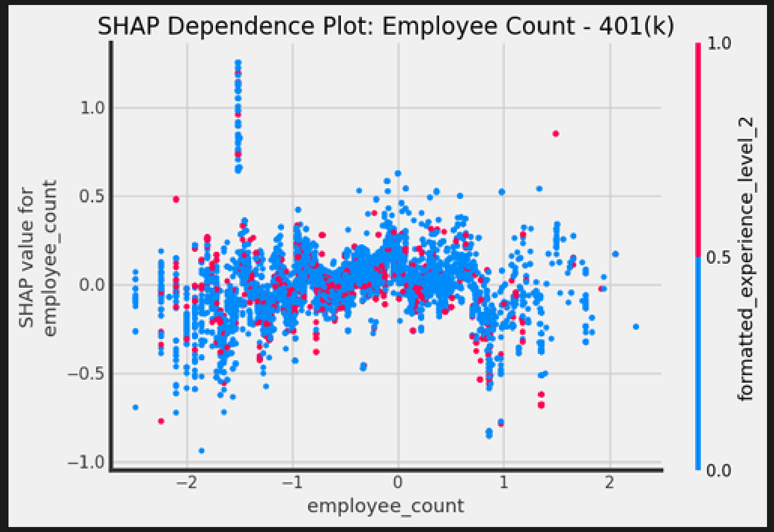
Estimation



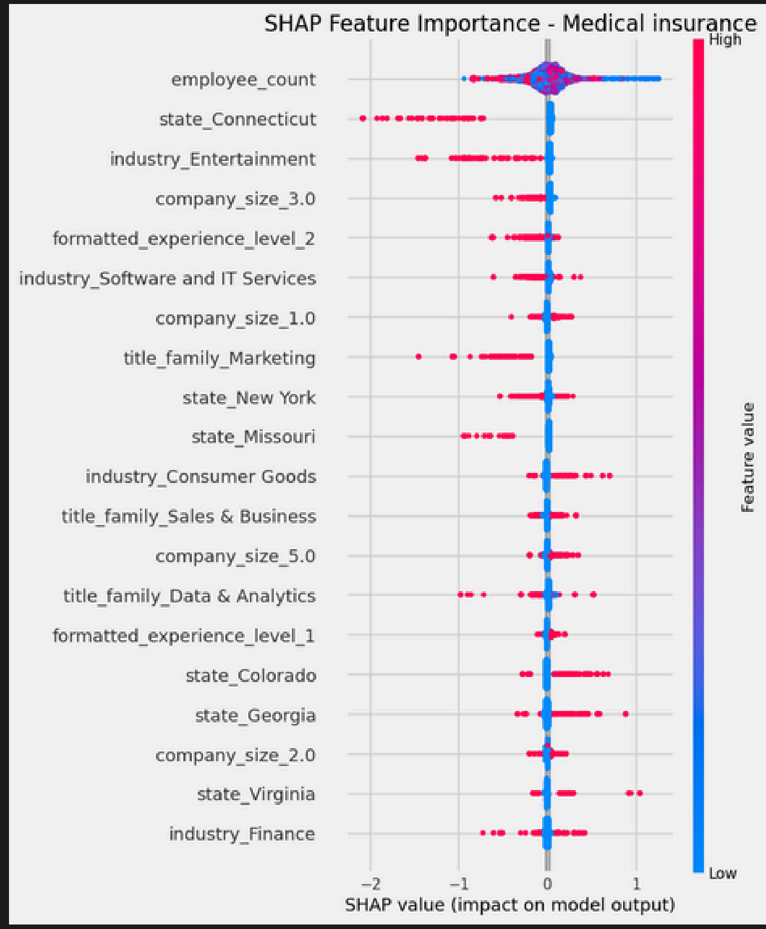


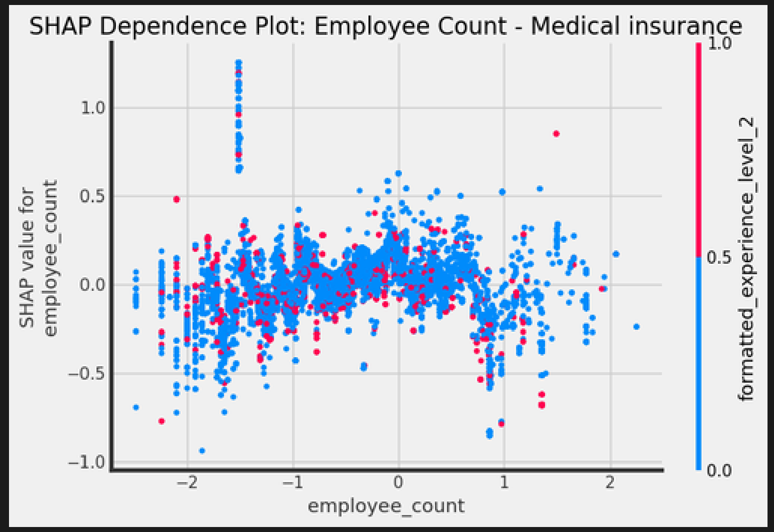
remote_allowed



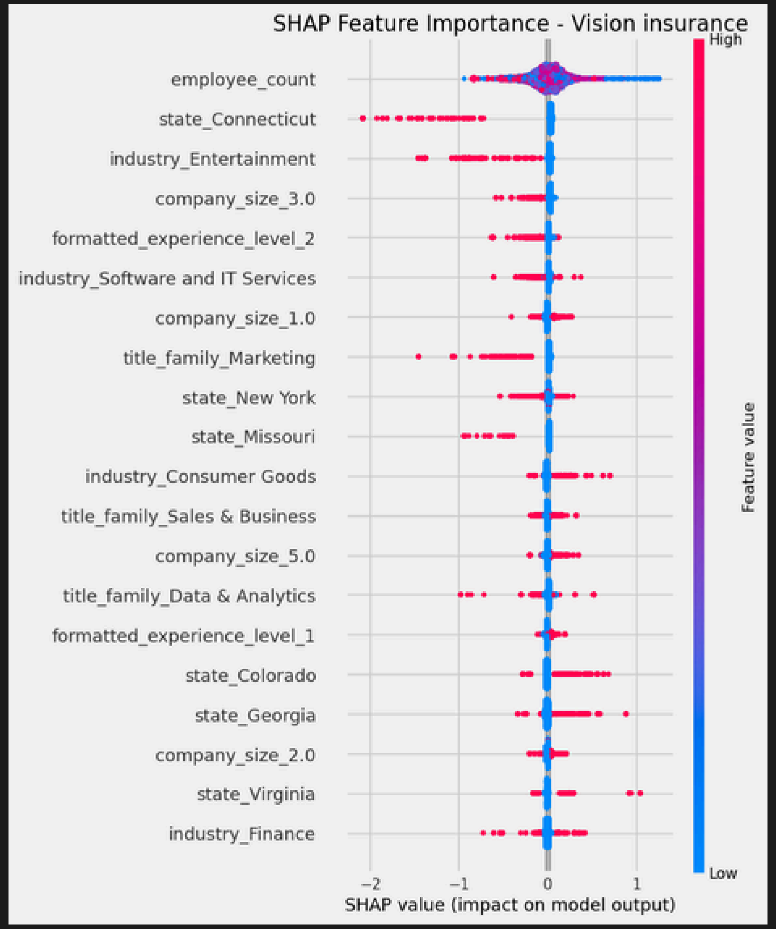


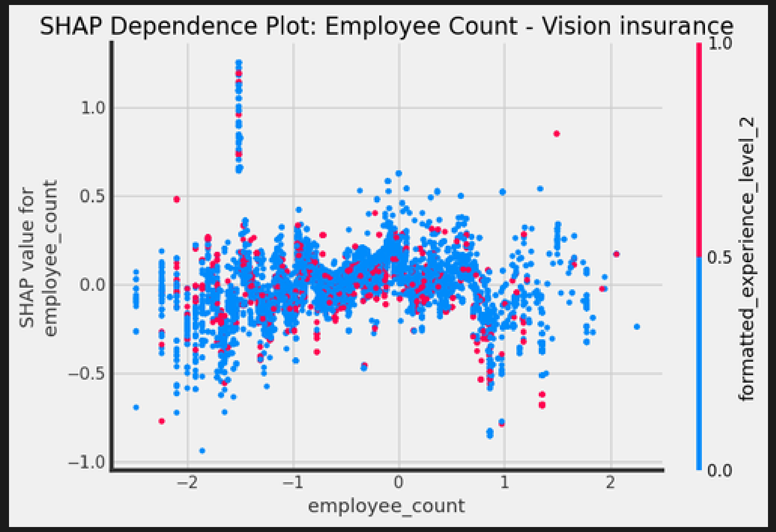
401 K





Medical indurance



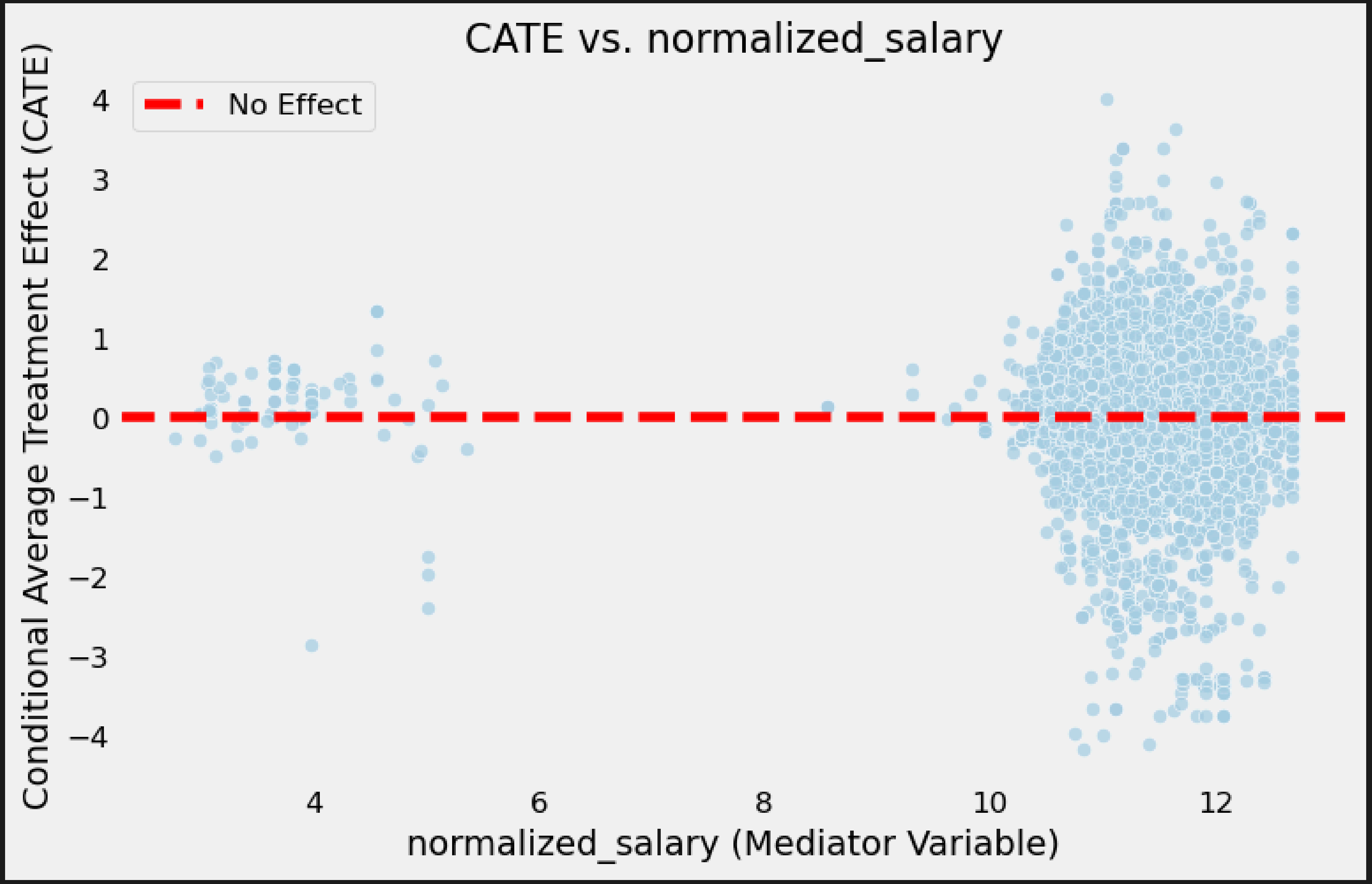


Vision indurance

CAUSAL INFERENCE

9

Estimation



For lower salaries (~3-6)

CATE varies more, with some negative effects --> offering benefits alone may not always compensate for lower salaries

For higher salaries (~10-12)

Data points cluster around CATE = 0 --> benefits have limited additional impact on engagement at high salary levels

CAUSAL INFERENCE

9

Estimation

Titles

```
# Data & Analytics
"data scientist": "Data & Analytics",
"data analyst": "Data & Analytics",
"data engineer": "Data & Analytics",
"business analyst": "Data & Analytics",
"financial analyst": "Data & Analytics",
"senior data": "Data & Analytics",
```

Industries

```
# Retail
"Retail": "Retail",
"Retail Groceries": "Retail",
"Wholesale": "Retail",
"Retail Apparel and Fashion": "Retail",
```

```
# Media and Communications
"Market Research": "Media and Communications",
"Marketing & Advertising": "Media and Communications",
"Newspaper Publishing": "Media and Communications",
"Online Audio and Video Media": "Media and Communications",
"Printing Services": "Media and Communications",
"Public Relations and Communications Services": "Media and Communications",
"Publishing": "Media and Communications",
"Translation and Localization": "Media and Communications",
"Writing and Editing": "Media and Communications",
"Advertising Services": "Media and Communications",
```

Heatmap of CATE by Industry & Title

Industry \ Job Title	Accounting & Finance	Customer Support	Data & Analytics	Engineering	Legal & Real Estate	Logistics & Supply Chain	Marketing	Medical & Healthcare	Operations & HR	Project/Product Management	Sales & Business	Senior Leadership	Software Engineering
Agriculture	0.34	-0.07								-0.11			
Construction	0.12	-0.19	-0.36	0.26		-0.12			0.41	0.06	-0.21	0.12	0.04
Consumer Goods	0.43	1.72	0.39	-0.03		-0.35	0.69		0.01	0.07	0.20	0.42	-0.67
Corporate Services	0.25	0.06	0.19	0.01	-0.01	0.06	0.24	0.09	-0.04	0.16	0.18	-0.04	-0.44
Design	0.48	0.20		0.47				0.49	0.56	-0.19	0.95		1.30
Education	0.76	0.03	-0.42	-0.11			-0.26	-0.13	-0.02	-0.11	0.12	-0.28	-0.27
Energy and Mining	-0.81	-0.31		0.05	0.28				-0.54	-0.46	-0.02		-0.16
Entertainment	-0.62	-0.03	-1.17		-0.39		-0.55	-1.66	-0.05	-0.86	-0.48	-0.11	-1.16
Finance	0.16	0.19	-0.65	0.14	-0.65		0.40	-0.13	0.05	0.21	0.39	0.31	0.02
Hardware and Networking	-0.31		-0.82	-0.31	0.50	-0.11	-0.57		-0.73	-0.21	-0.36	-0.22	-0.17
Healthcare	-0.04	0.04	0.12	-0.19			0.30	-0.01	0.02	0.03	0.36	0.09	-0.05
Legal	0.10	0.53	-0.18		-0.17		0.36	-0.06	0.33	-0.21	0.54	-0.04	-0.40
Manufacturing	0.09	-0.02	-1.66	0.21		0.37	0.44	2.03	0.17	0.09	0.06	0.77	-0.21
Media and Communications	0.27	0.72	1.17	0.08			0.15		-1.35	-0.42	0.15	0.59	1.05
Nonprofit	0.31	0.14	0.27	1.70			-0.76	0.10	0.24	0.29	0.44	0.35	
Public Administration	0.32	-0.06	-0.11	-0.43	-0.24			-0.08	0.36	0.10	-0.13	0.34	0.25
Public Safety	0.32		-0.17	0.18					0.40	-0.19	0.12		0.56
Real Estate	0.64	-0.04	-0.24		0.06		0.39		0.15	0.17	0.58	0.89	-1.02
Recreation and Travel	-0.02	0.27	-1.43	0.38			0.42	-0.00	-0.03	0.15	-0.14	0.08	-0.03
Retail	-0.01	-0.02	-0.39			0.19	0.20		0.83	0.17	0.32	0.29	1.01
Software and IT Services	0.12	0.16	-0.25	0.45	0.13	-0.03	0.14	0.01	0.02	0.33	0.01	-0.26	0.19
Transportation and Logistics	0.19	0.27	-0.48			0.06	0.65		0.05	0.14	0.14	0.32	-0.73
Wellness and Fitness	-0.05	-0.12	-0.60	0.01		-2.66	0.32	0.08	0.21	-0.09	-0.07	-1.00	

Industry

Job Title

2 1 0 −1 −2

9 CAUSAL INFERENCE

Estimation

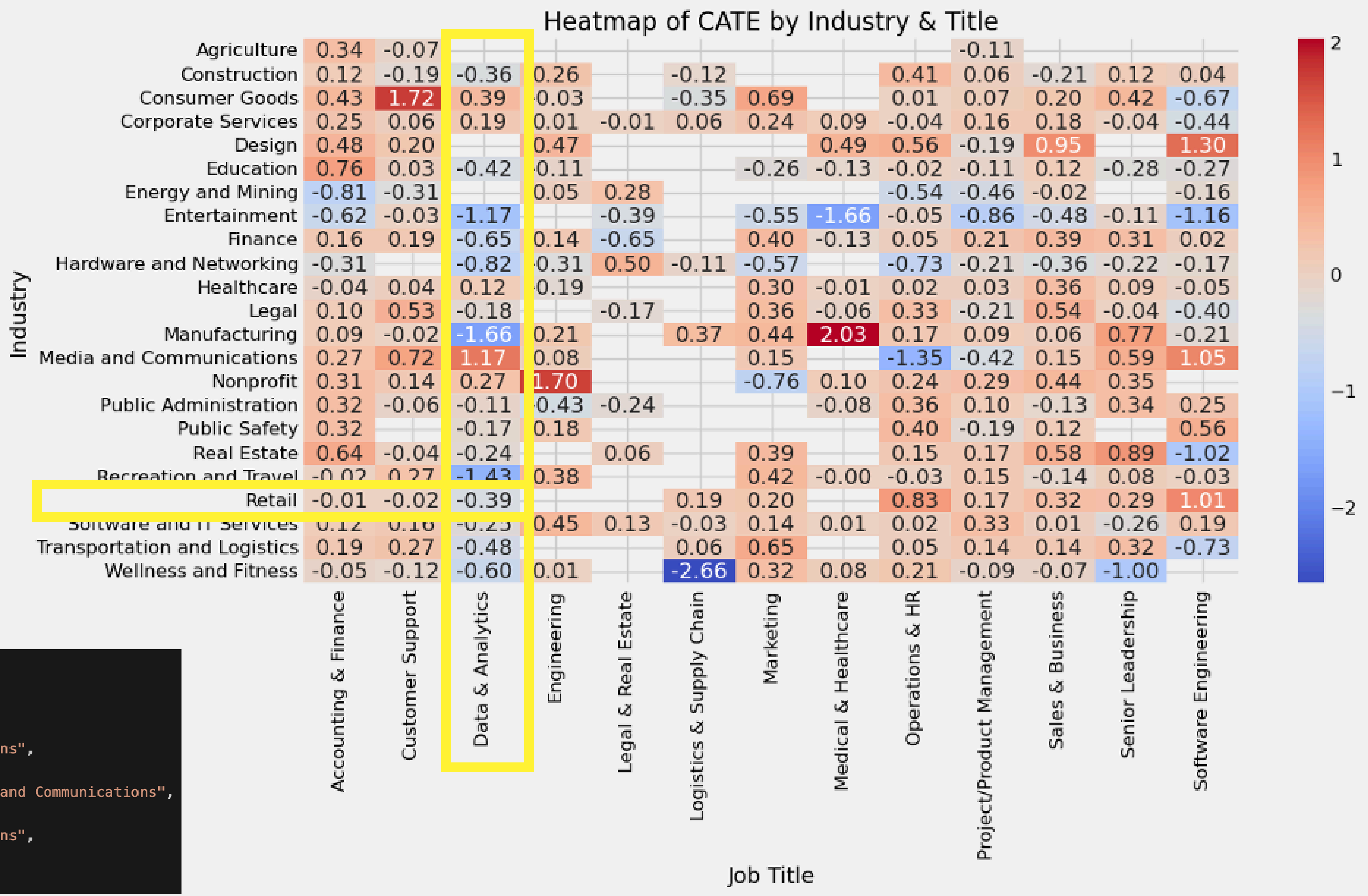
Titles

```
# Data & Analytics
"data scientist": "Data & Analytics",
"data analyst": "Data & Analytics",
"data engineer": "Data & Analytics",
"business analyst": "Data & Analytics",
"financial analyst": "Data & Analytics",
"senior data": "Data & Analytics",
```

Industries

```
# Retail
"Retail": "Retail",
"Retail Groceries": "Retail",
"Wholesale": "Retail",
"Retail Apparel and Fashion": "Retail",
```

```
# Media and Communications
"Market Research": "Media and Communications",
"Marketing & Advertising": "Media and Communications",
"Newspaper Publishing": "Media and Communications",
"Online Audio and Video Media": "Media and Communications",
"Printing Services": "Media and Communications",
"Public Relations and Communications Services": "Media and Communications",
"Publishing": "Media and Communications",
"Translation and Localization": "Media and Communications",
"Writing and Editing": "Media and Communications",
"Advertising Services": "Media and Communications",
```



CAUSAL INFERENCE

Refutation Tests

Refutation Summary Across Benefits:

Benefit	Original Effect	Random Common Cause	Placebo Effect	Subset Effect	Random Change (%)	Placebo Change (%)	Subset Change (%)
remote_allowed	0.6908030920549518	0.6907947013641201	0.002062953517294121	0.6915748722884923	0.0012146284416206627	99.70136880668015	0.11172217415018845
401(k)	0.027687242295725906	0.02767092087464515	0.0010187562073829517	0.027892284390984505	0.05894924783923428	96.32048509381444	0.7405652504809087
Medical insurance	0.11216525674546696	0.11215481802973785	0.0016424242181714735	0.11351699369977795	0.009306550024479268	98.53571037429302	1.2051298178530017
Vision insurance	0.12462325054081744	0.12462576441594309	0.0017549306864286121	0.12537411264767337	0.0020171798719251	98.59181117583366	0.6025056348614535

CAUSAL INFERENCE

(Experience Level Model - Multi-level treatment) Failed Experiment

confounders

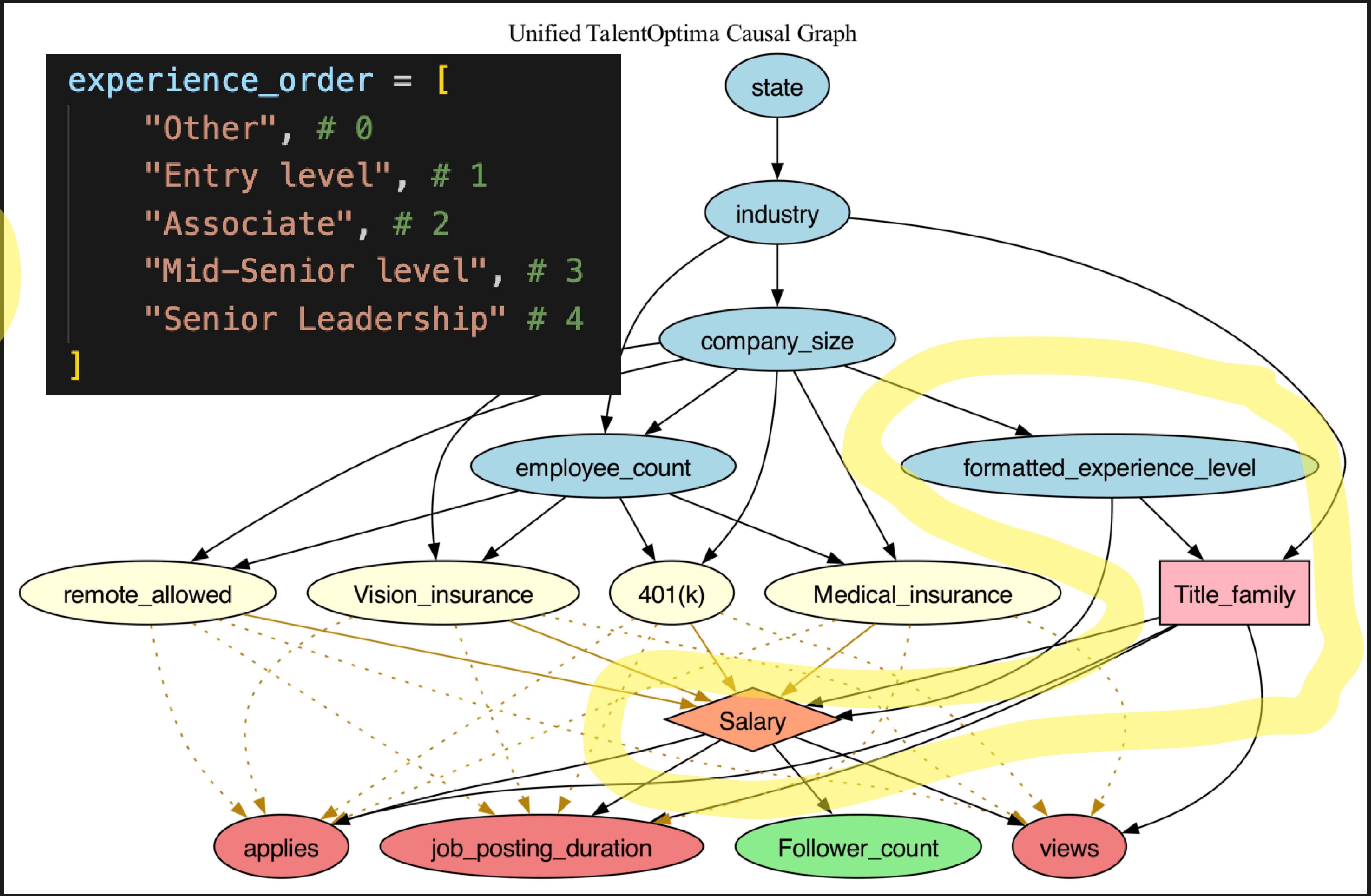
Treatments (Intervention)

Treatments (Intervention)

mediator

Post-treatment factors (backdoor > 1)

Post-treatment factors (backdoor = 1)



CAUSAL INFERENCE

(Experience Level Model - Multi-level treatment) Failed Experiment

SMD + GPS

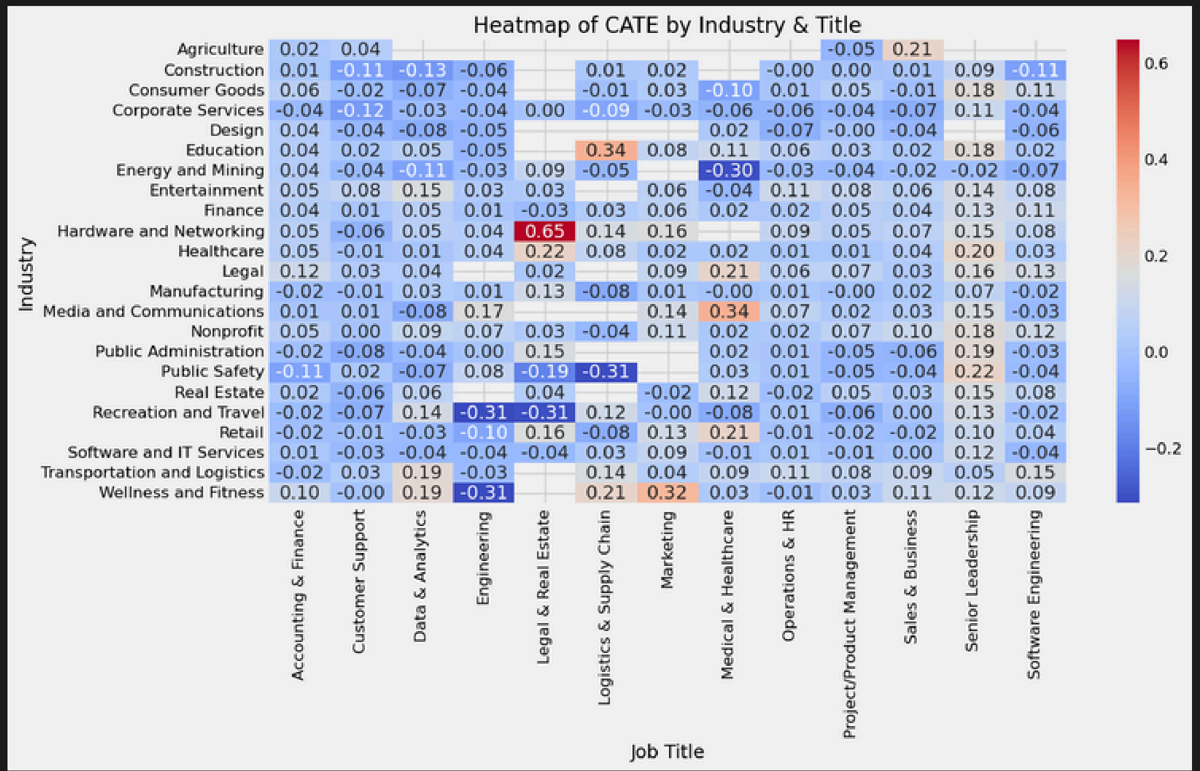
```
SMD Before Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 215
Large imbalance (SMD > 0.25): 41

SMD After Matching Summary:
Extreme imbalance (SMD > 1): 0
Some imbalance (SMD > 0.1): 20
Large imbalance (SMD > 0.25): 4

Matched Sample Sizes per Treatment Level:
Treatment Level 0: 6045 samples
Treatment Level 1: 6045 samples
Treatment Level 2: 0 samples
Treatment Level 3: 0 samples
Treatment Level 4: 0 samples
```

OLS (ATE)

```
Mediation Analysis Results:
Total Effect of Experience on Salary:
const                          -0.232030
formatted_experience_level_1   -0.196768
dtype: float64
Indirect Effect via Title: -0.0029463682201514682
Direct Effect of Experience on Salary:
formatted_experience_level_1   -0.052137
dtype: float64
Total Effect (Direct + Indirect):
formatted_experience_level_1   -0.055084
dtype: float64
```

RLearner + XGBoost (CATE)

```
Estimated Heterogeneous Treatment Effects (HTE) per Treatment Level:
formatted_experience_level
0   -0.022112
1   -0.125633
2   -0.104301
3    0.064679
4    0.291328
Name: tau_hat, dtype: float64
```



Robustness Tests

```
Refutation using Random Common Cause:
Refute: Add a random common cause
Estimated effect:0.07315654101107721
New effect:0.07315635756224806
p value:0.98
```

```
Refutation using Data Subset:
Refute: Use a subset of data
Estimated effect:0.07315654101107721
New effect:0.07318060120476405
p value:1.0
```

```
Refutation using Placebo Treatment:
Refute: Use a Placebo Treatment
Estimated effect:0.07315654101107721
New effect:3.1402564460144155e-05
p value:0.98
```

Thank You

Website www.TalentOptima.com

E-mail Invest@TalentOptima.com

Phone +123-456-7890/+123-456-7890

HQ address 123 Sherbrook St., Montreal, QC 12345