



IT Salaries 3.0

Lelia Erscoi

Agenda



About the Data

KNN - K-Fold

LDA - K-Fold

Logistic Regression - K-Fold

KNN - Leveling Gender Entries

LDA- Leveling Gender Entries

Logistic Regression - Leveling Gender Entries

Feature Importance

Findings

About the Data

Object Columns

	Gender	Position	Seniority	MainSkill	SecondarySkills	VacationDays	Employment	ContractDuration	CompanySize	CompanyType	LossOfJobCorona
0	Male	Software Engineer	Senior	TypeScript	Kotlin, Javascript / Typescript	30	Full-time employee	Unlimited contract	51-100	Product	No
1	Male	Backend Developer	Senior	Ruby		NaN	28	Full-time employee	Unlimited contract	101-1000	Product
2	Male	Frontend Developer	Junior	Javascript		NaN	24	Full-time employee	Unlimited contract	51-100	Startup
3	Male	Backend Developer	Senior	C# .NET	.NET, SQL, AWS, Docker	29	Full-time employee	Unlimited contract	101-1000	Product	No
4	Male	DevOps	Senior	AWS, GCP, Python,K8s	Python, AWS, Google Cloud, Kubernetes, Docker	30	Full-time employee	Unlimited contract	11-50	Startup	No

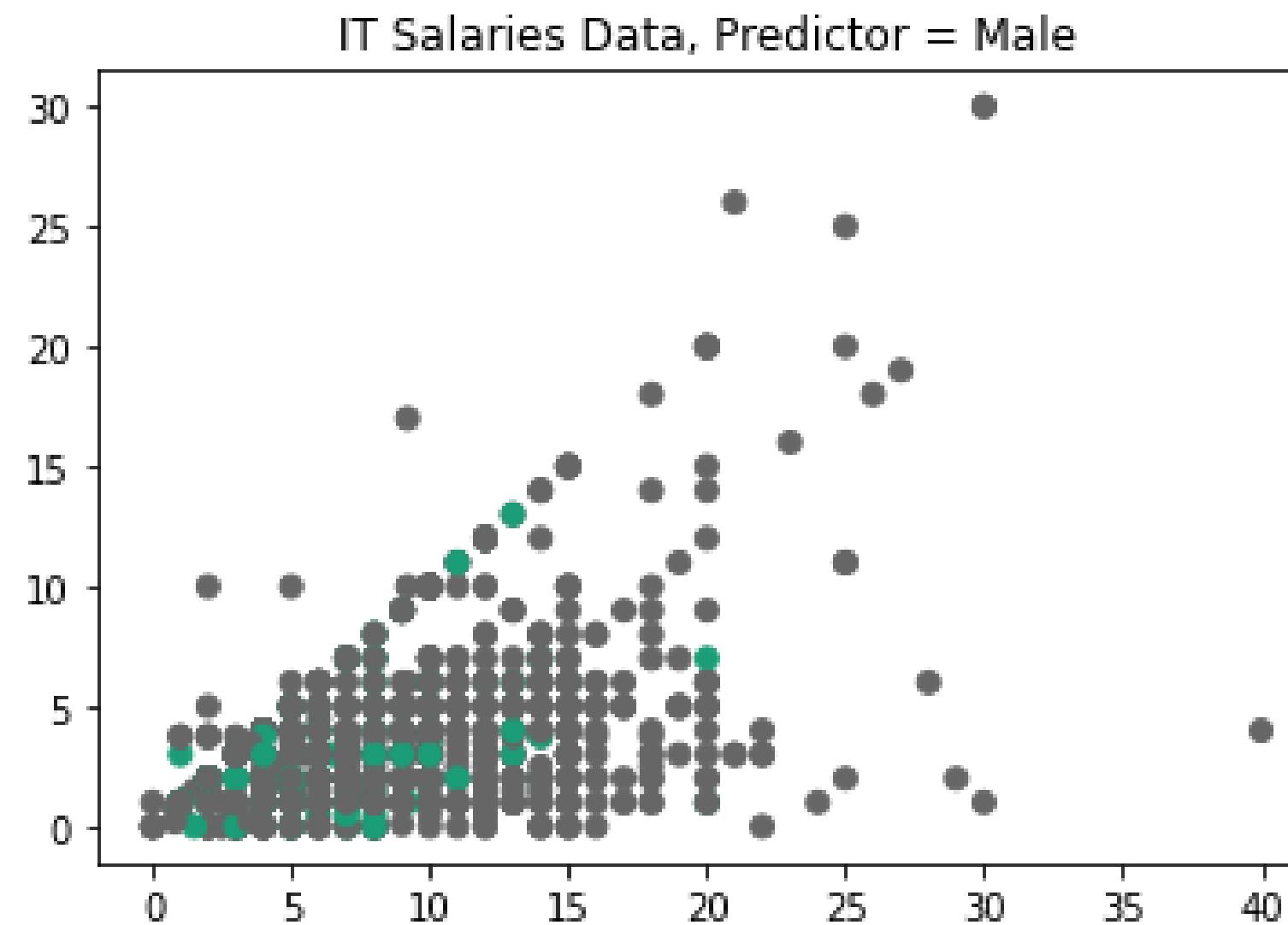
About the Data

No More Object Columns!

Age	Position	TotalYearsExperience	YearsExperienceGermany	MainSkill	SecondarySkills	BrutSalary	BrutSalaryAndStocks	AnnualBrutLastYear	AnnualBrutAndStocksLastYear
31.000000	71777.000000	11.0	1.0	67670.905109	68772.489878	75000.0	5000.0	65000.0	5000.0
39.000000	71777.000000	19.0	7.0	80819.861250	69636.731123	87000.0	97000.0	82000.0	92000.0
38.000000	71777.000000	12.0	6.0	70612.933333	76197.651996	92000.0	40000.0	85000.0	30000.0
31.000000	69636.731123	10.0	2.0	67670.905109	65895.882353	80000.0	0.0	78000.0	5000.0
32.568935	66232.417582	4.0	4.0	66380.952388	78079.651734	105000.0	0.0	85000.0	0.0
33.000000	69636.731123	10.0	1.5	69636.731123	65895.882353	60000.0	3000.0	48000.0	2500.0

Encoding

About the Data

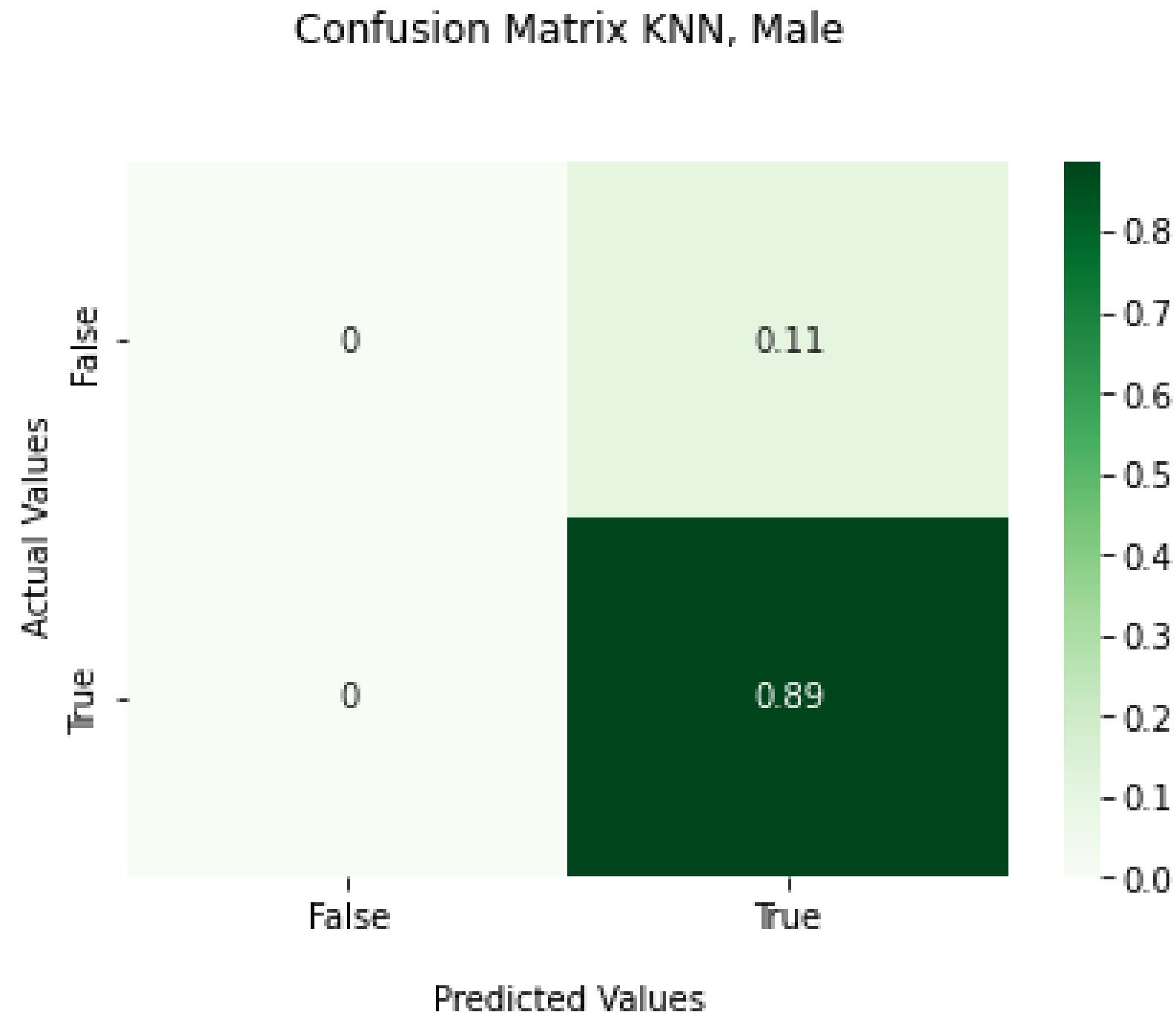


Insights:

- 1031 x 21 data points
- Focus on the Gender dimension
- 141 "female" and 890 "male"

KNN

Predicting Male Label



```
Avg accuracy : 0.8682242990654206
precision      recall     f1-score    support
          0         0.00      0.00      0.00       12
          1         0.89      1.00      0.94      94
accuracy
macro avg       0.44      0.50      0.47      106
weighted avg    0.79      0.89      0.83      106

Minimum error:- 0.15942028985507245 at K = 6
Minimum error:- 0.13592233009708737 at K = 8
Minimum error:- 0.15048543689320387 at K = 8
Minimum error:- 0.14563106796116504 at K = 2
Minimum error:- 0.1407766990291262 at K = 7
```

KNN

Predicting Male Label

Before:

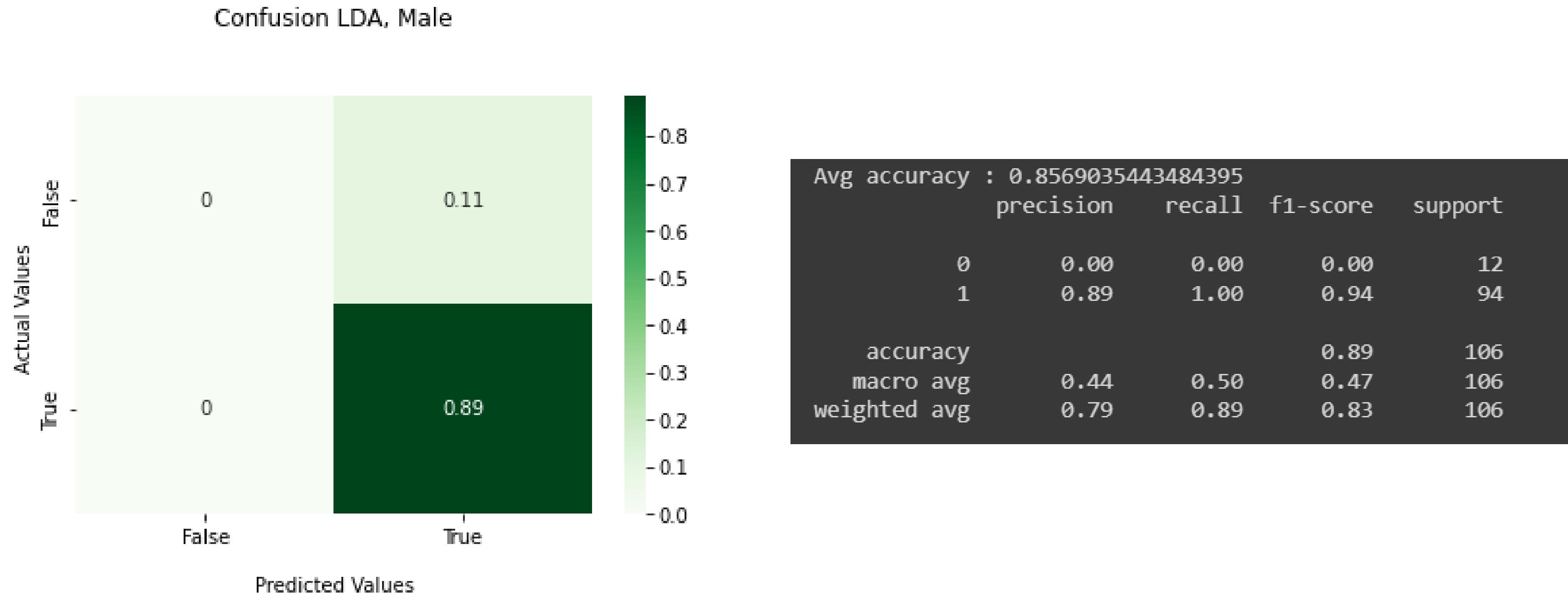
Accuracy
87.5%

After:

Metric/ Fold	Best K	Accuracy
1	6	84.11%
2	2	83.39%
3	6	89.62%
4	4	87.73%
5	4	88.67%

LDA

Predicting Male Label



LDA

Predicting Male Label

Before:

Accuracy
86.69%

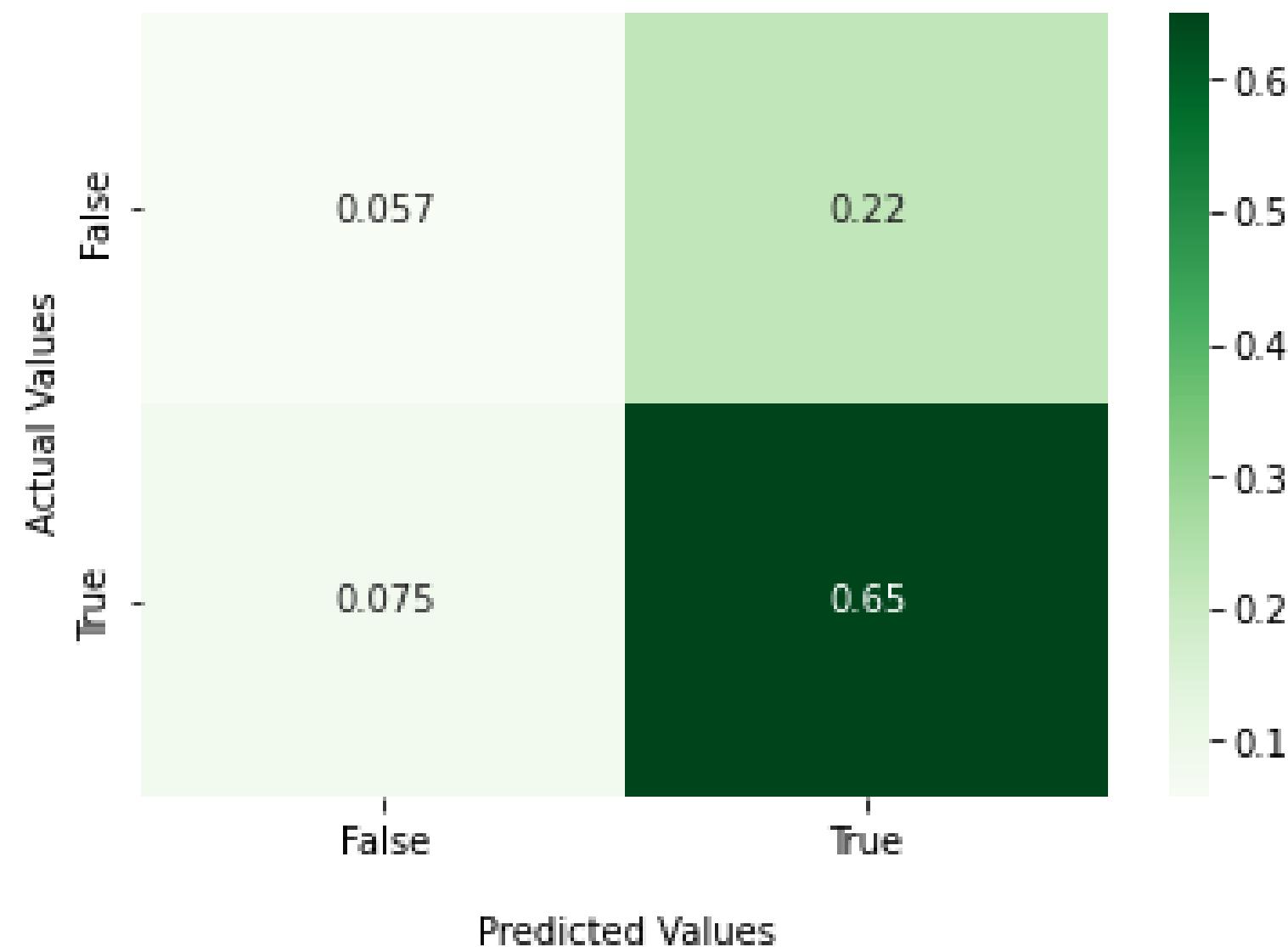
After:

Metric/ Fold	Accuracy
1	84.11%
2	84.9%
3	88.67%
4	82.07%
5	88.67%

Logistic Regression

Predicting Male Label

Confusion Matrix Logistic Regression, Male



Avg accuracy : 0.7382472227120438					
	precision	recall	f1-score	support	
0	0.43	0.21	0.28	29	
1	0.75	0.90	0.82	77	
accuracy			0.71	106	
macro avg	0.59	0.55	0.55	106	
weighted avg	0.66	0.71	0.67	106	

Logistic Regression

Predicting Male Label

Before:

Accuracy
86.69%

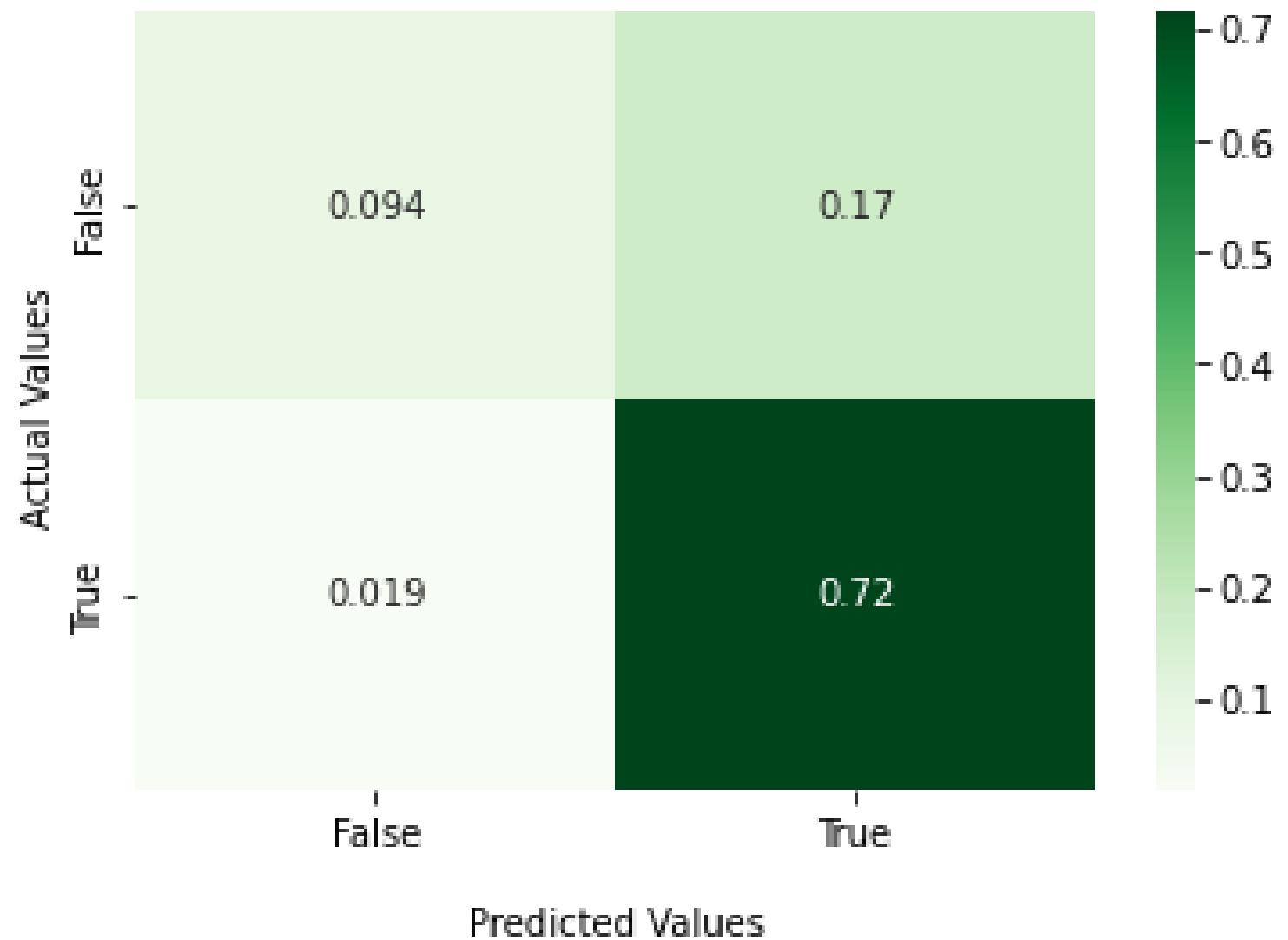
After:

Metric/ Fold	Accuracy
1	72.89%
2	75.47%
3	75.47%
4	74.52%
5	70.75%

K Nearest Neighbors

Results Leveling Entries

Confusion Matrix KNN, Male Label with Fewer Men

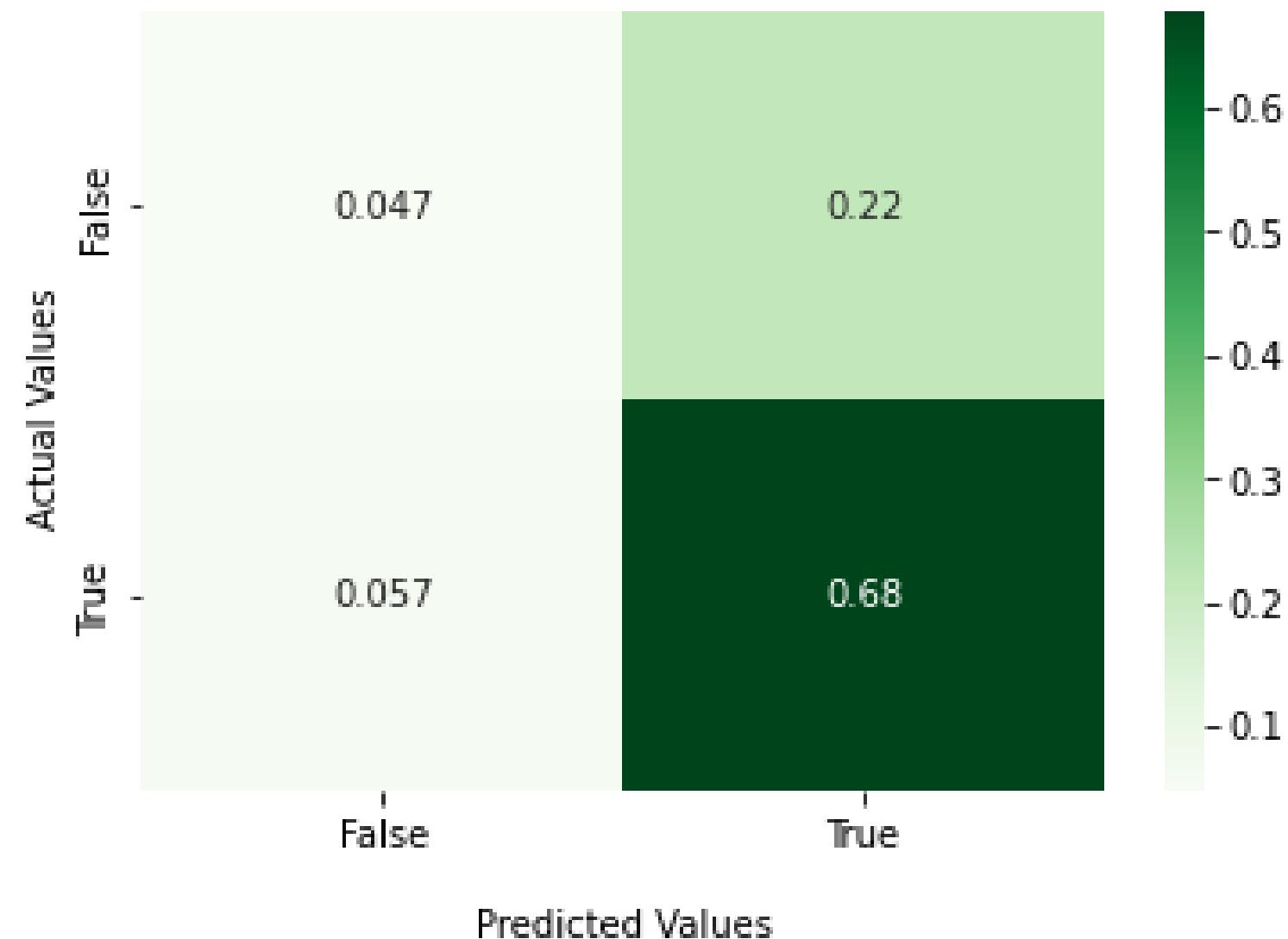


	precision	recall	f1-score	support
0	0.83	0.36	0.50	28
1	0.81	0.97	0.88	78
accuracy			0.81	106
macro avg	0.82	0.67	0.69	106
weighted avg	0.82	0.81	0.78	106

LDA

Results Leveling Entries

Confusion Matrix LDA, Male Label with Fewer Men

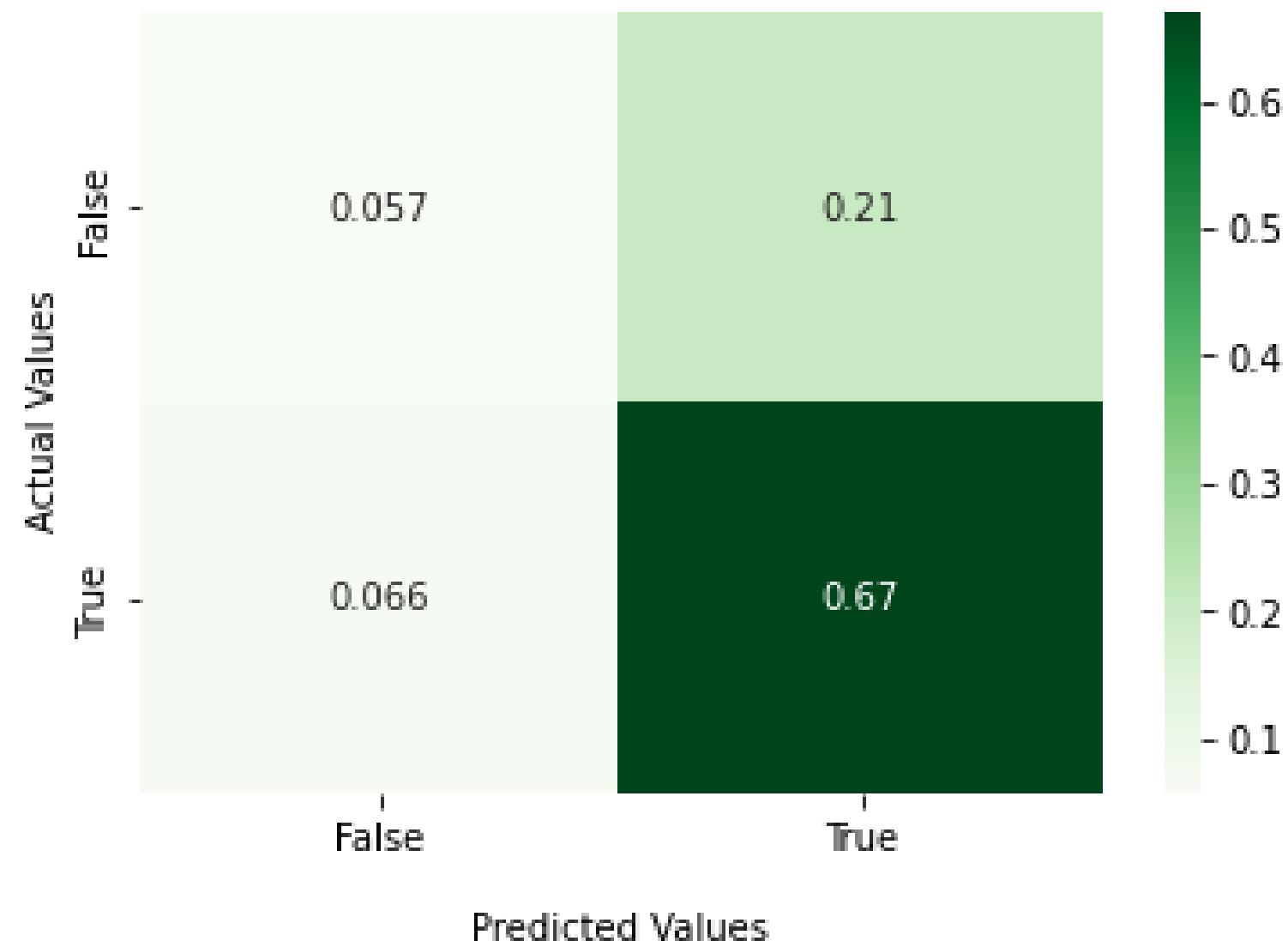


	precision	recall	f1-score	support
0	0.45	0.18	0.26	28
1	0.76	0.92	0.83	78
accuracy			0.73	106
macro avg	0.61	0.55	0.54	106
weighted avg	0.68	0.73	0.68	106

Logistic Regression

Results Leveling Entries

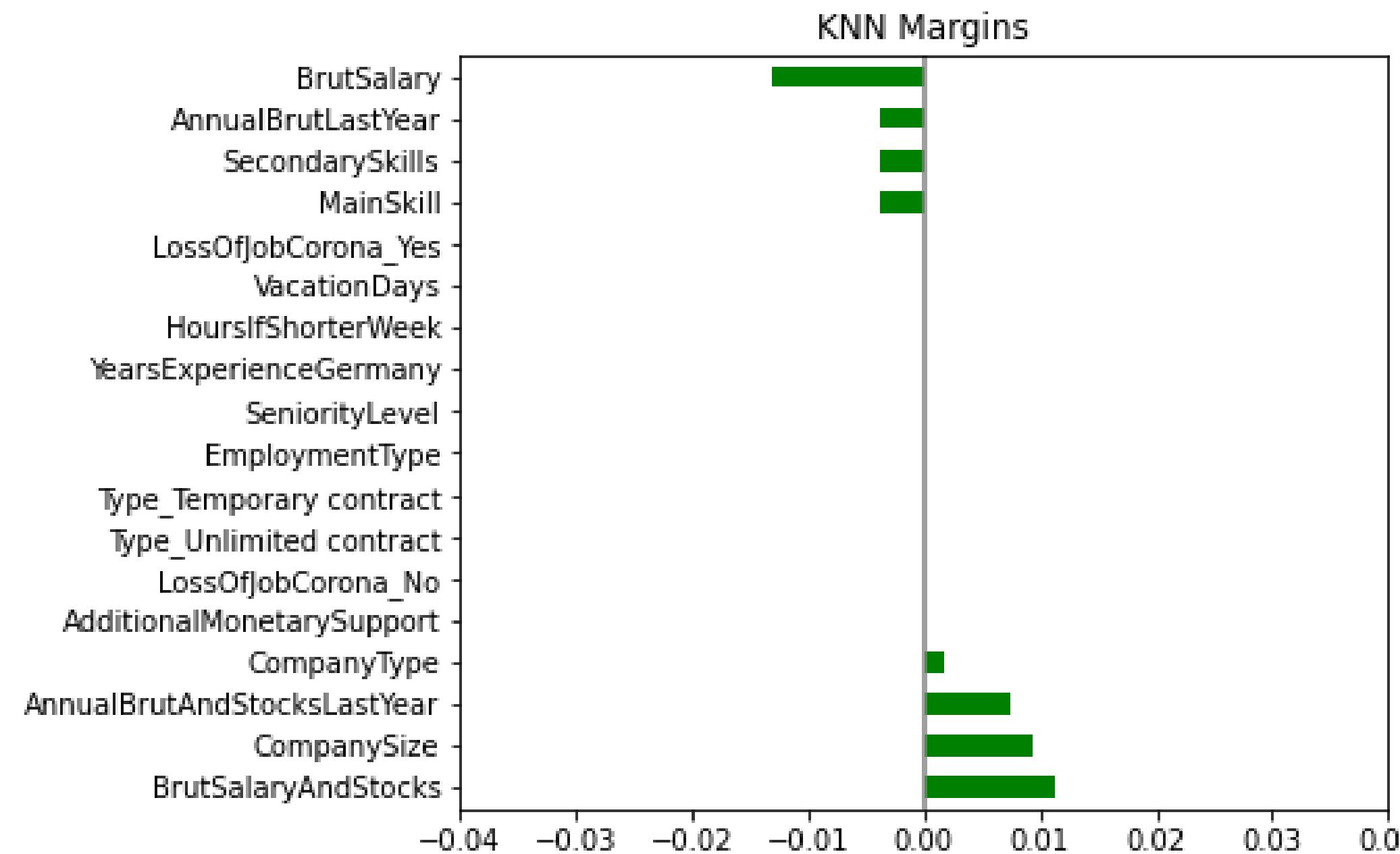
Confusion Matrix Logistic Regression, Male Label with Fewer Men



	precision	recall	f1-score	support
0	0.46	0.21	0.29	28
1	0.76	0.91	0.83	78
accuracy			0.73	106
macro avg	0.61	0.56	0.56	106
weighted avg	0.68	0.73	0.69	106

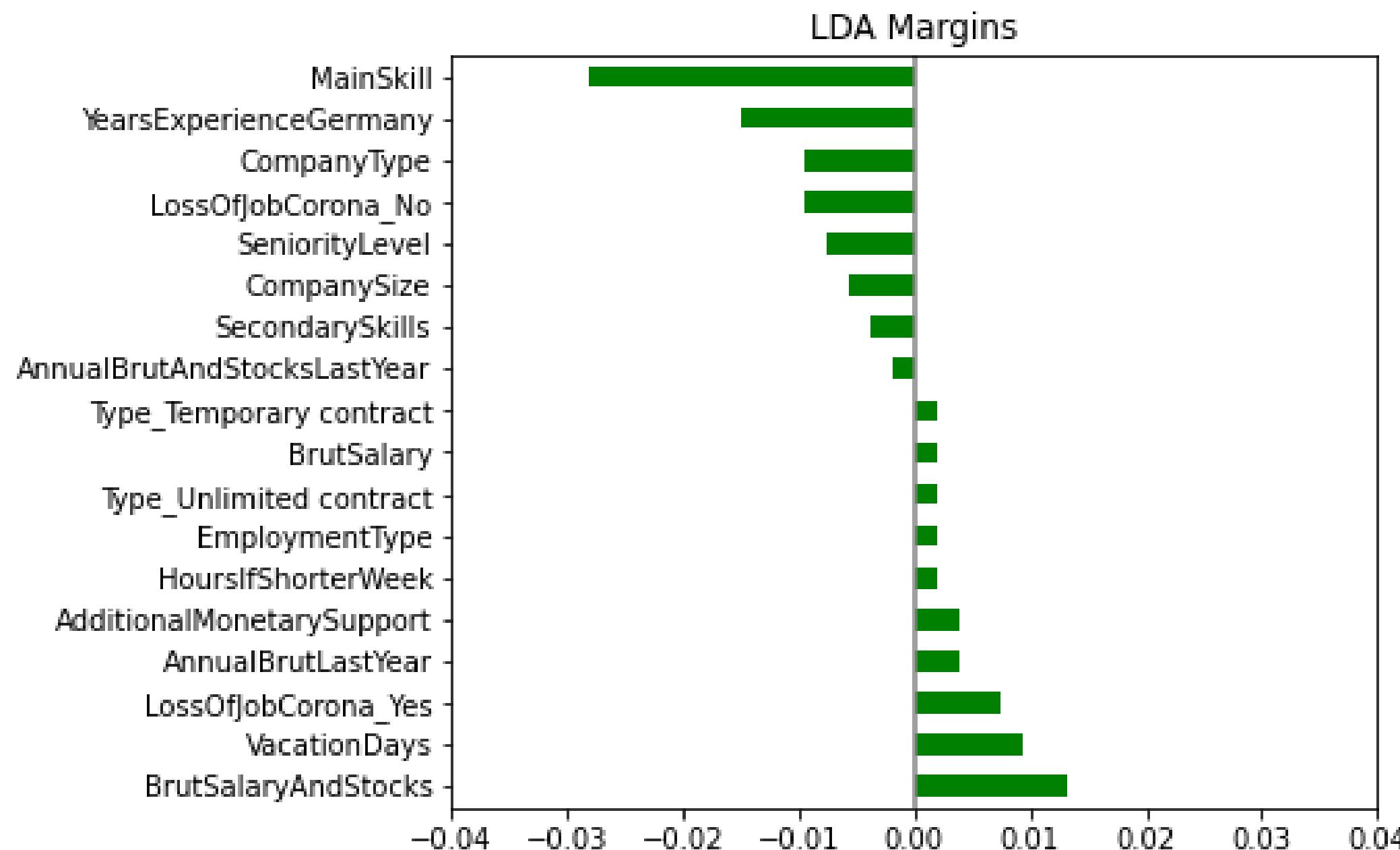
KNN

Feature Importance



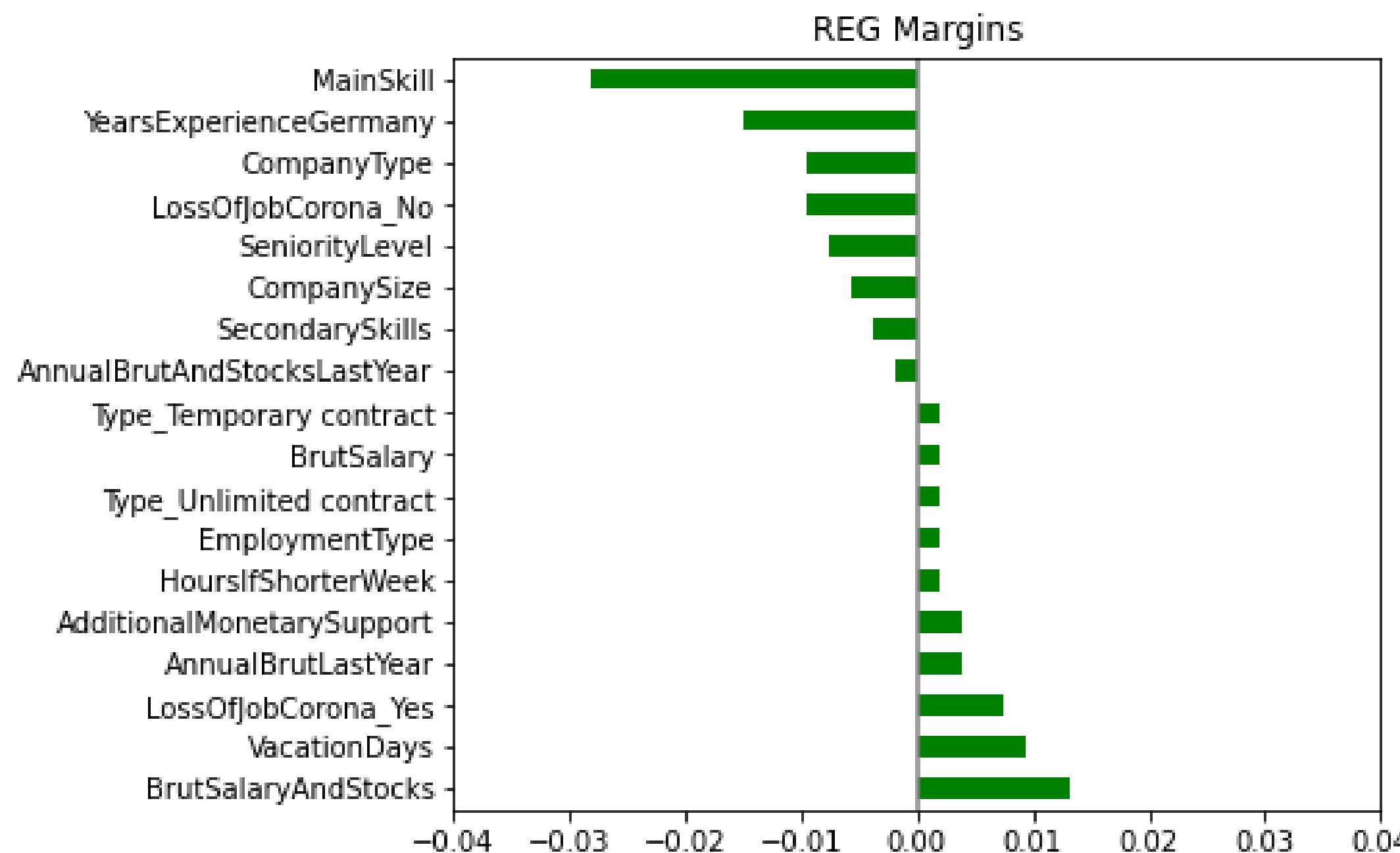
LDA

Feature Importance

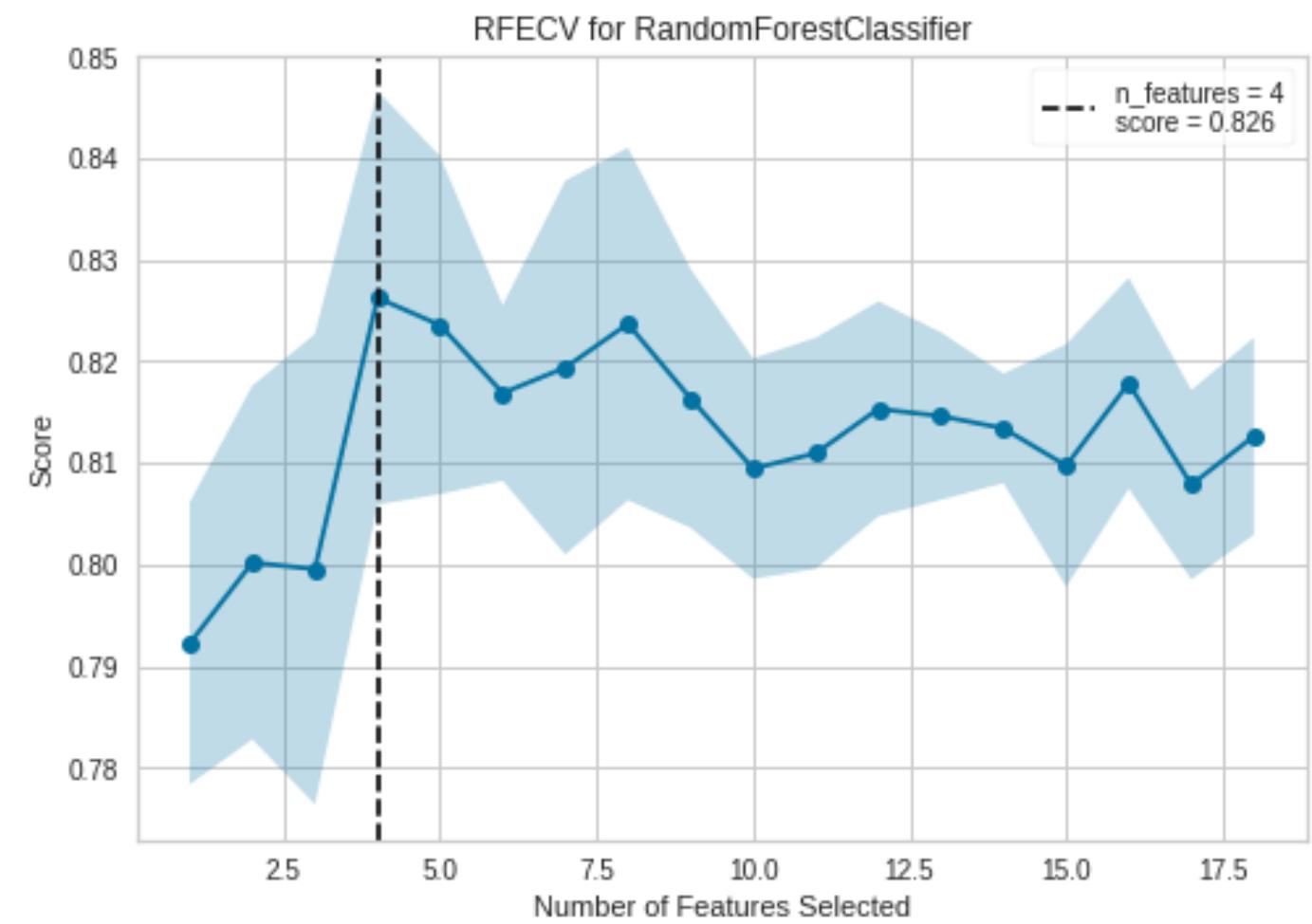
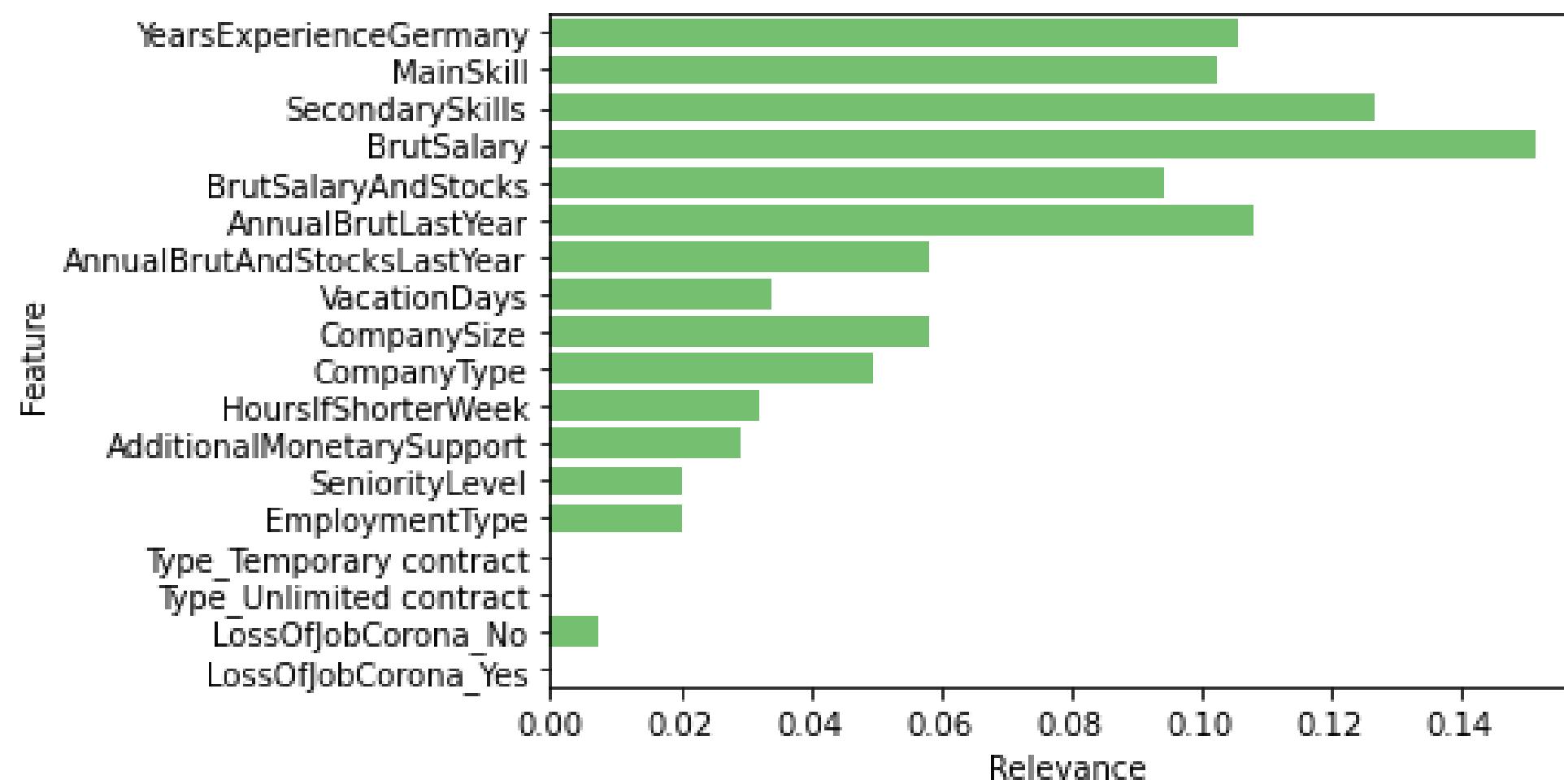


Logistic Regression

Feature Importance



Findings



Thank You!

