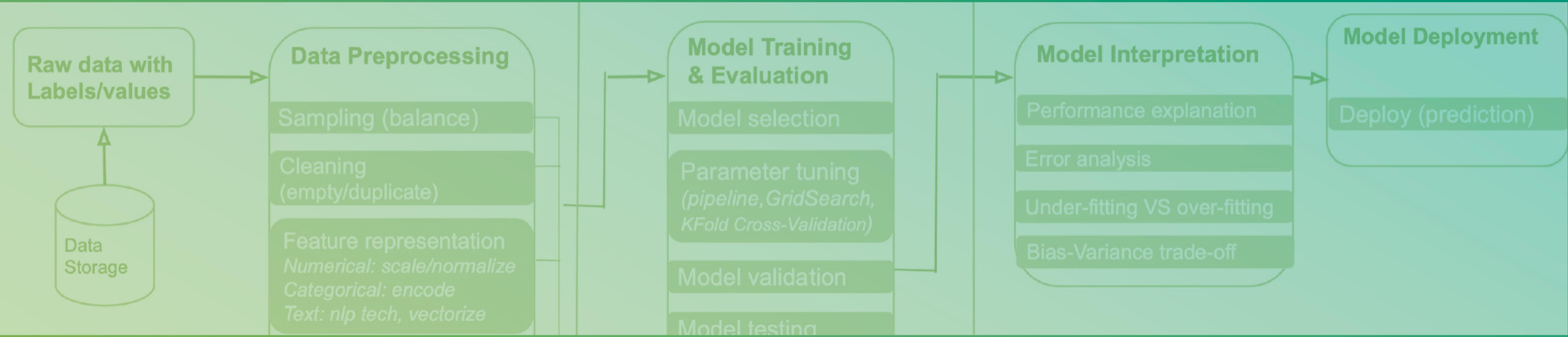
FACTORS IN INFLUENCING JOB CHANGE

By Senling Shu



GOALS:

01

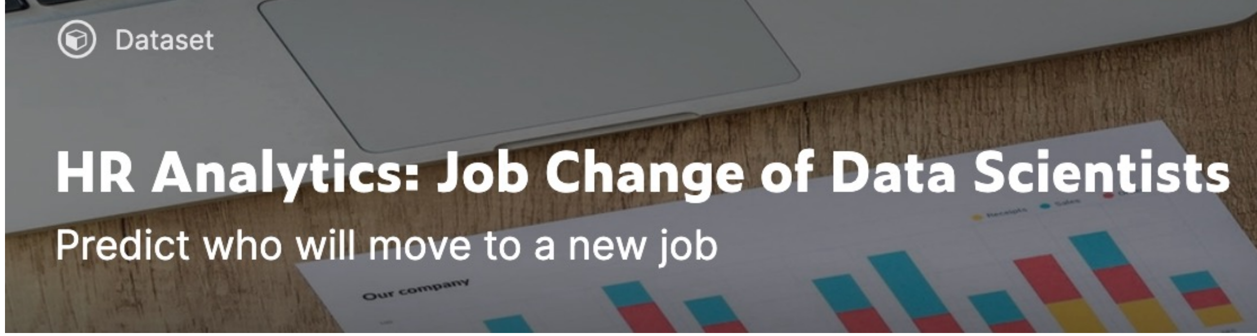
To understand what kinds of factors are most predictive of data scientists leaving their current job

=> Classification

02

To accomplish the whole process of building a machine learning pipeline

Data



Dataset

HR Analytics: Job Change of Data Scientists

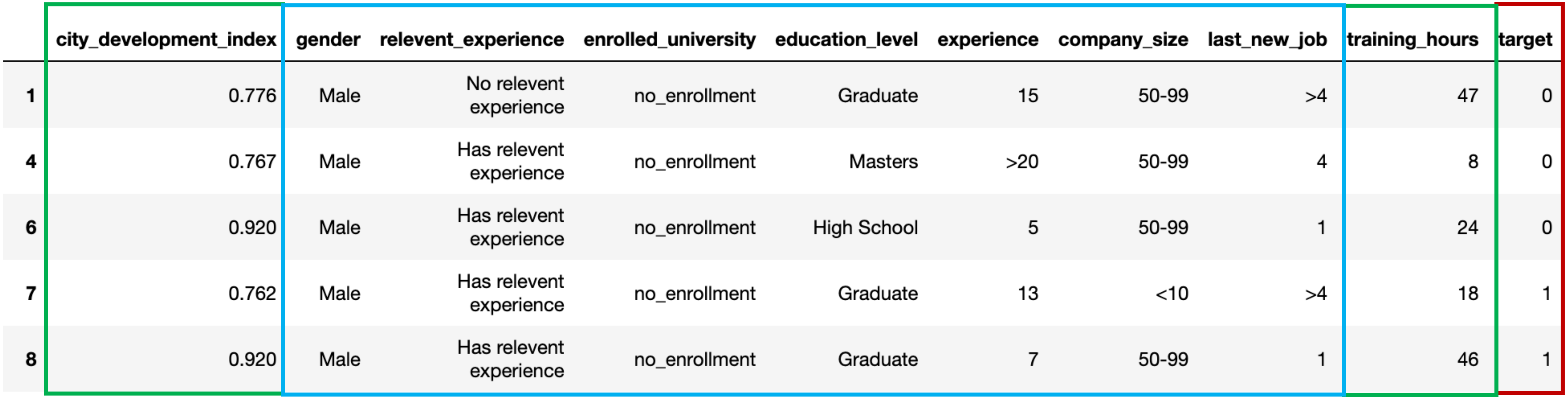
Predict who will move to a new job

Predictors: 9 features (**2 numerical**, **7 categorical**)

Target:

0 – looking for a job change (14,381 samples)

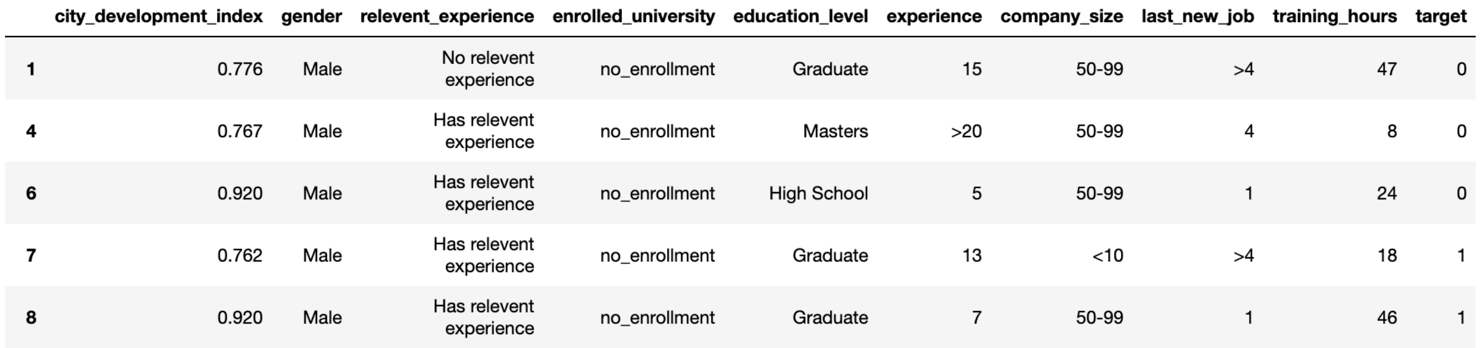
1 – not looking for a job change (4,777 samples)



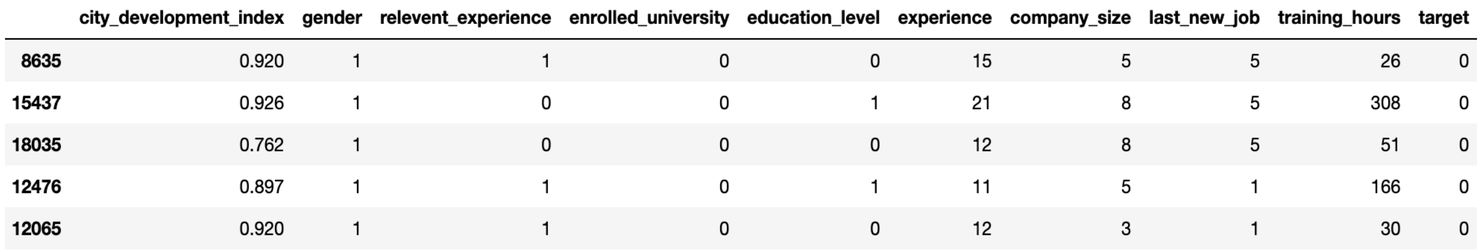
	city_development_index	gender	relevent_experience	enrolled_university	education_level	experience	company_size	last_new_job	training_hours	target
1	0.776	Male	No relevent experience	no_enrollment	Graduate	15	50-99	>4	47	0
4	0.767	Male	Has relevent experience	no_enrollment	Masters	>20	50-99	4	8	0
6	0.920	Male	Has relevent experience	no_enrollment	High School	5	50-99	1	24	0
7	0.762	Male	Has relevent experience	no_enrollment	Graduate	13	<10	>4	18	1
8	0.920	Male	Has relevent experience	no_enrollment	Graduate	7	50-99	1	46	1

Data Preprocessing

1. **Drop NA** ➔ 19158 to 10129 samples (0: 8501 1: 1628)
2. **Sampling** ➔ balance the 0s and 1s ➔ randomly select 1628 samples from the 0 class (reduced the total sample size to 3,256)
3. **Encoding** ➔ encode all 7 categorical features to numerical values
4. **Scaling + Normalization** ➔ normalized with both l1 & l2 norm
5. **Train/Test Split** with a test size of 0.2



	city_development_index	gender	relevent_experience	enrolled_university	education_level	experience	company_size	last_new_job	training_hours	target
1	0.776	Male	No relevent experience	no_enrollment	Graduate	15	50-99	>4	47	0
4	0.767	Male	Has relevent experience	no_enrollment	Masters	>20	50-99	4	8	0
6	0.920	Male	Has relevent experience	no_enrollment	High School	5	50-99	1	24	0
7	0.762	Male	Has relevent experience	no_enrollment	Graduate	13	<10	>4	18	1
8	0.920	Male	Has relevent experience	no_enrollment	Graduate	7	50-99	1	46	1



	city_development_index	gender	relevent_experience	enrolled_university	education_level	experience	company_size	last_new_job	training_hours	target
8635	0.920	1	1	0	0	15	5	5	26	0
15437	0.926	1	0	0	1	21	8	5	308	0
18035	0.762	1	0	0	0	12	8	5	51	0
12476	0.897	1	1	0	1	11	5	1	166	0
12065	0.920	1	1	0	0	12	3	1	30	0

Model Selection & Parameter Tuning

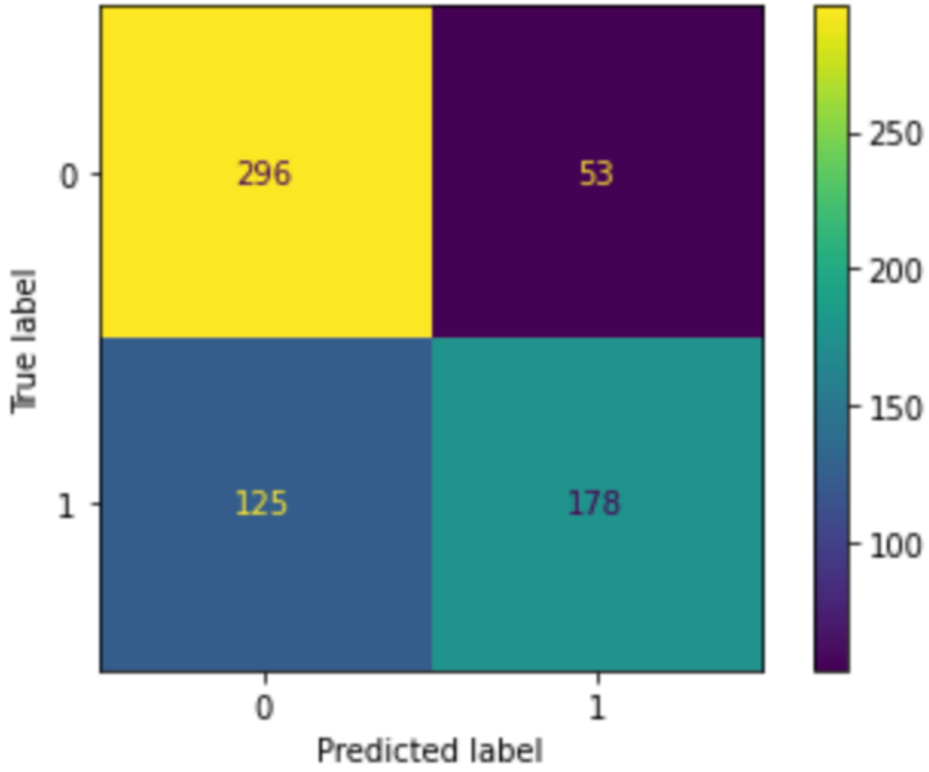
- SVC: {'C':[0.1, 1, 10, 100,1000]} with a Linear Kernel
- Decision Tree: {'criterion':('gini', 'entropy'), 'max_depth':[3,4,5]}
- Logistic Regression: {'C':[0.1, 1, 10, 100, 1000], 'penalty': ['l1','l2','none']}

All searched with a **5**-fold cross-validation & **l1** + **l2** normalization !

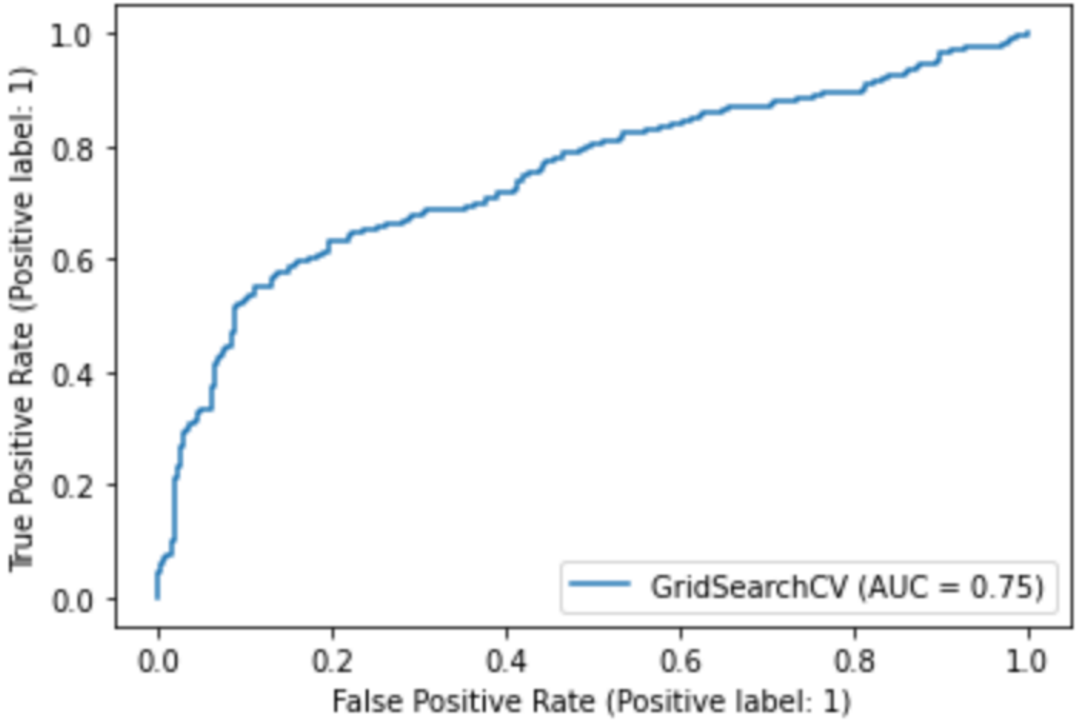
SVC

Best Parameters: {'C': 1} with l2 norm

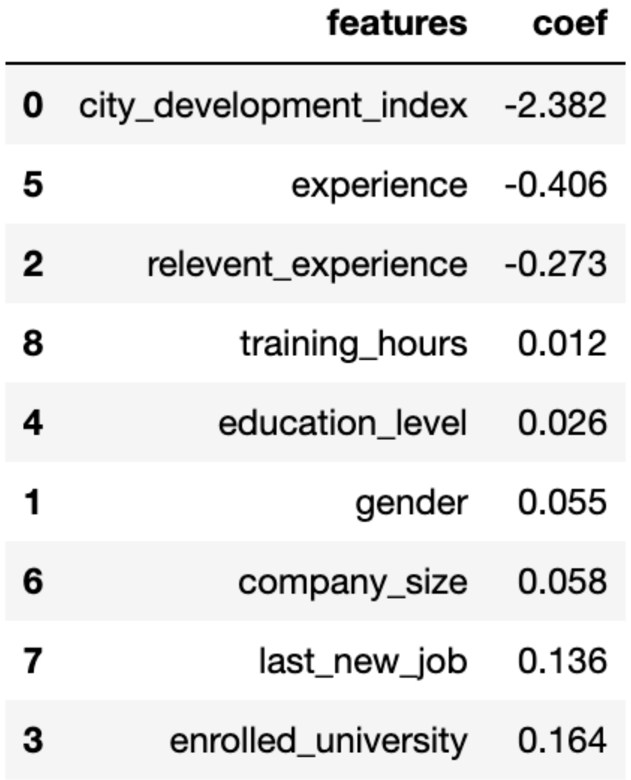
Metrics	Values
Accuracy	0.727
Precision (for 1)	0.734
Precision (for 0)	0.703
F1	0.721
Recall	0.727



tpr = 0.587



fpr = 0.152



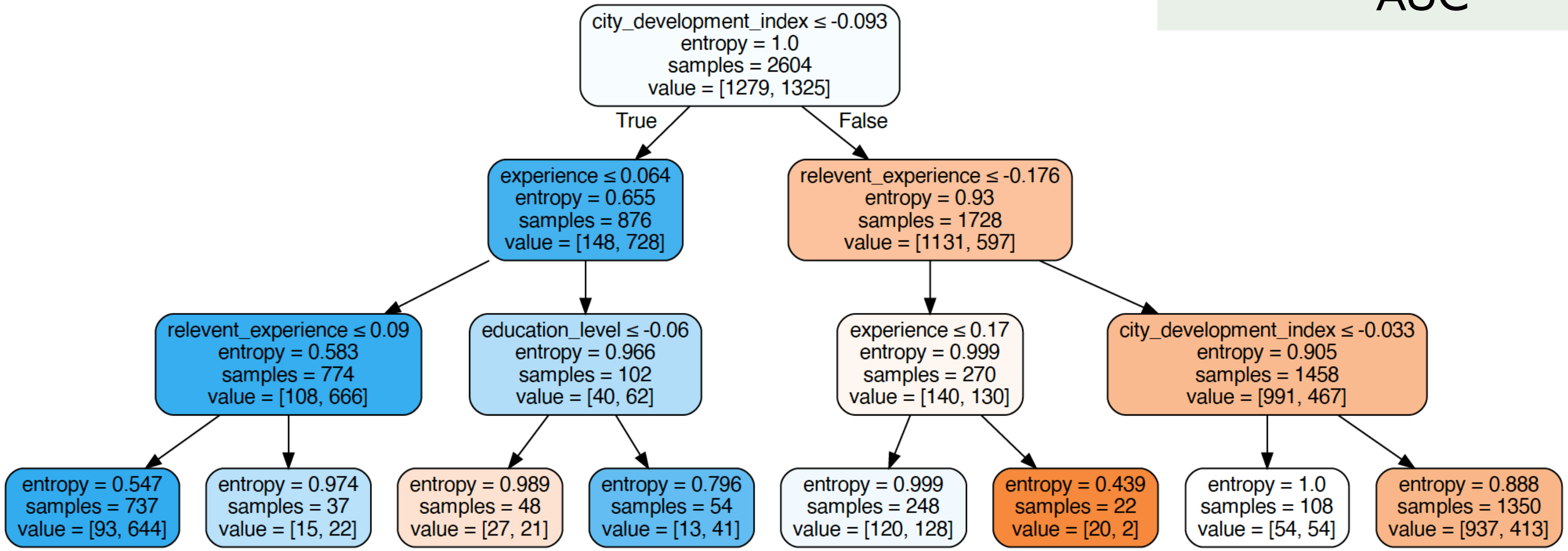
	features	coef
0	city_development_index	-2.382
5	experience	-0.406
2	relevent_experience	-0.273
8	training_hours	0.012
4	education_level	0.026
1	gender	0.055
6	company_size	0.058
7	last_new_job	0.136
3	enrolled_university	0.164

No enrollment,
part time enrollment,
Full time enrollment

Decision Tree

Best Parameters: {'criterion': 'entropy', 'max_depth': 3} with l1 norm

Metrics	Values
Accuracy	0.704
Precision (for 1)	0.706
Precision (for 0)	0.694
FPR	0.200
TPR	0.594
AUC	0.730



city_development_index ≤ -0.093
entropy = 1.0
samples = 2604
value = [1279, 1325]

True

False

experience ≤ 0.064
entropy = 0.655
samples = 876
value = [148, 728]

relevent_experience ≤ -0.176
entropy = 0.93
samples = 1728
value = [1131, 597]

relevent_experience ≤ 0.09
entropy = 0.583
samples = 774
value = [108, 666]

education_level ≤ -0.06
entropy = 0.966
samples = 102
value = [40, 62]

experience ≤ 0.17
entropy = 0.999
samples = 270
value = [140, 130]

city_development_index ≤ -0.033
entropy = 0.905
samples = 1458
value = [991, 467]

entropy = 0.547
samples = 737
value = [93, 644]

entropy = 0.974
samples = 37
value = [15, 22]

entropy = 0.989
samples = 48
value = [27, 21]

entropy = 0.796
samples = 54
value = [13, 41]

entropy = 0.999
samples = 248
value = [120, 128]

entropy = 0.439
samples = 22
value = [20, 2]

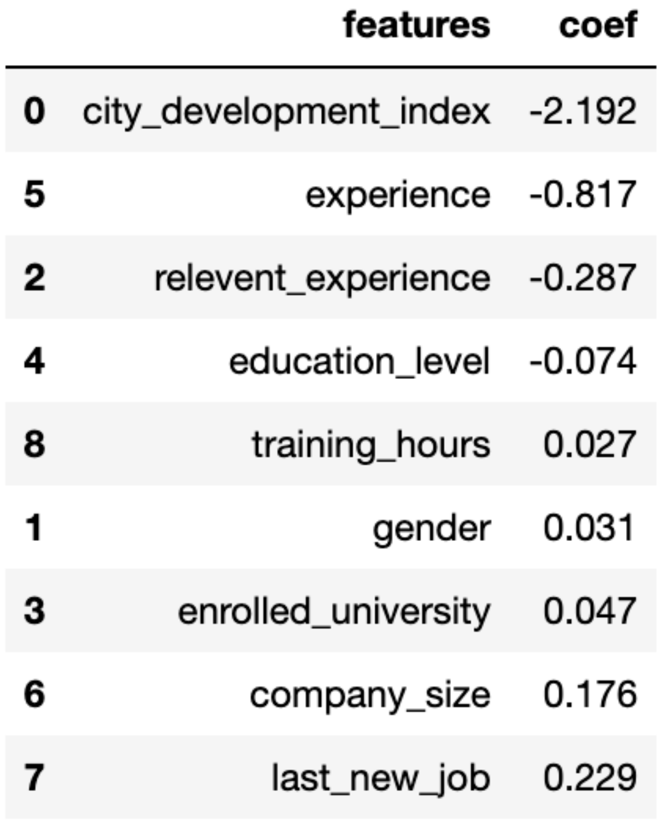
entropy = 1.0
samples = 108
value = [54, 54]

entropy = 0.888
samples = 1350
value = [937, 413]

Logistic Regression

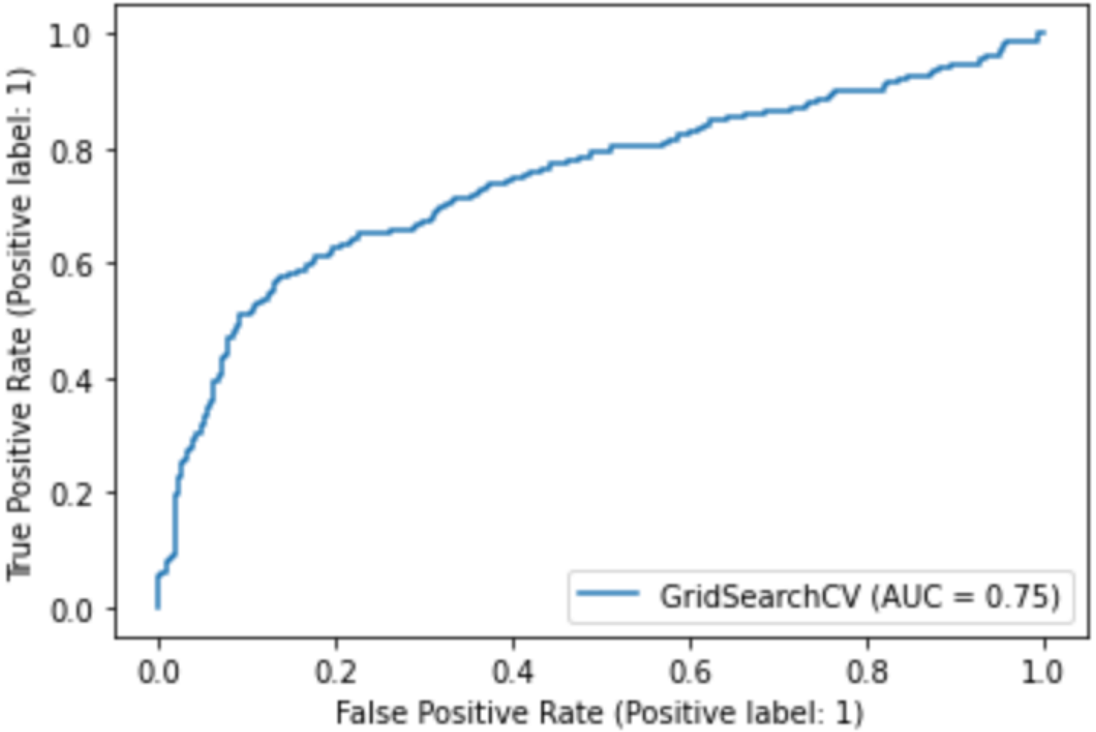
Best Parameters: {'C': 0.1, 'penalty': 'l2’} with l2 norm

Metrics	Values
Accuracy	0.716
Precision (for 1)	0.717
Precision (for 0)	0.714



	features	coef
0	city_development_index	-2.192
5	experience	-0.817
2	relevent_experience	-0.287
4	education_level	-0.074
8	training_hours	0.027
1	gender	0.031
3	enrolled_university	0.047
6	company_size	0.176
7	last_new_job	0.229

tpr = 0.637



fpr = 0.215

Conclusion

Model	Accuracy
SVC	0.727 ★
Decision Tree	0.704
Logistic Regression	0.716

The development of city matters!