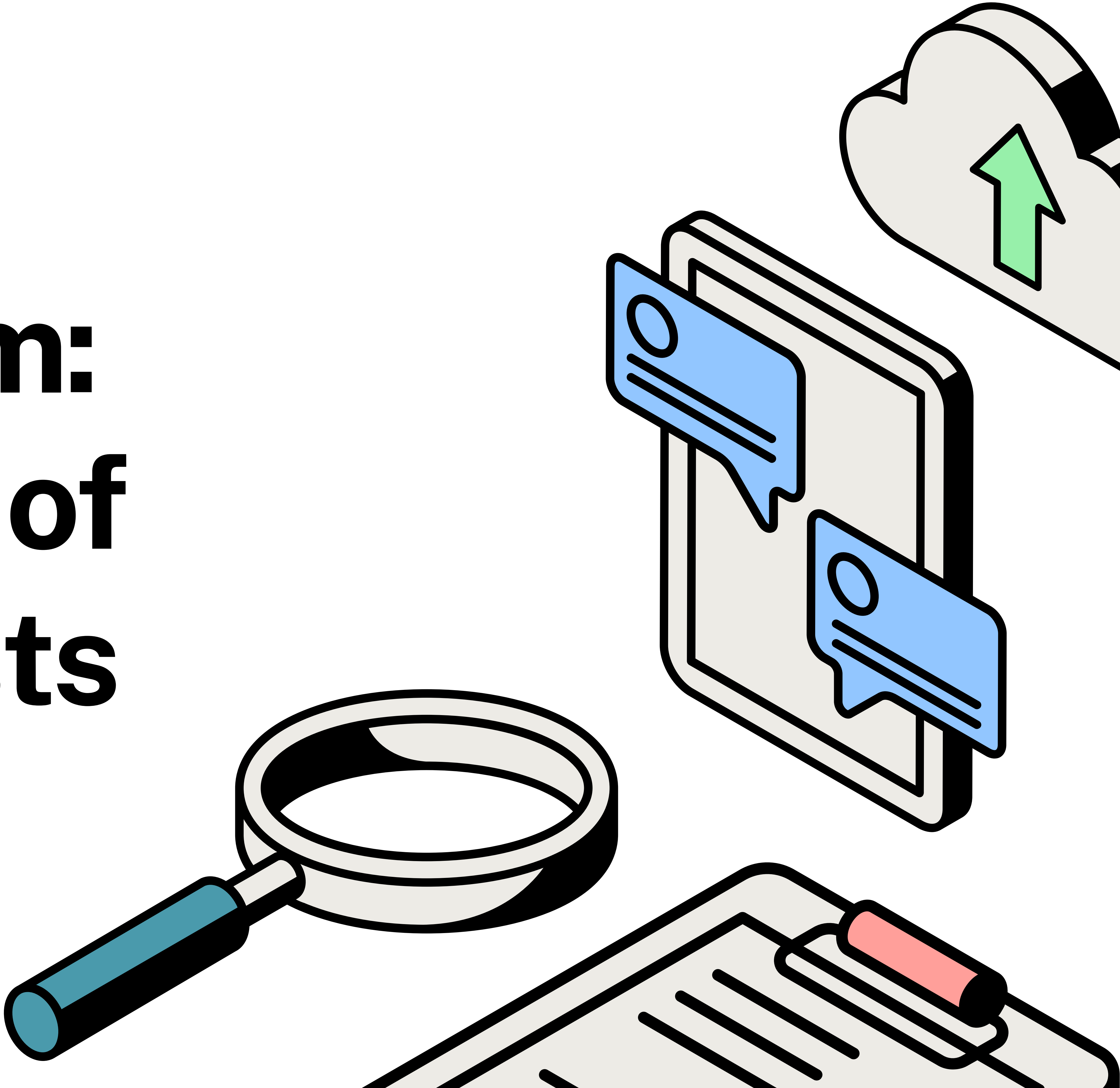
SIGMOID

Sigmoid Exam: HR Analytics of Data Scientists



Prepared by: Daniela Vornic

Examined by: Eduard Balamatiuc

Overview

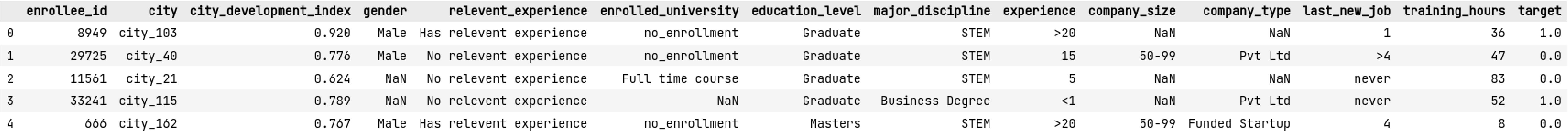
Context - HR Analytics

- A company in Big Data and Data Science wants to hire data scientists from successful course participants.
- Demographic, education, and experience data are available.

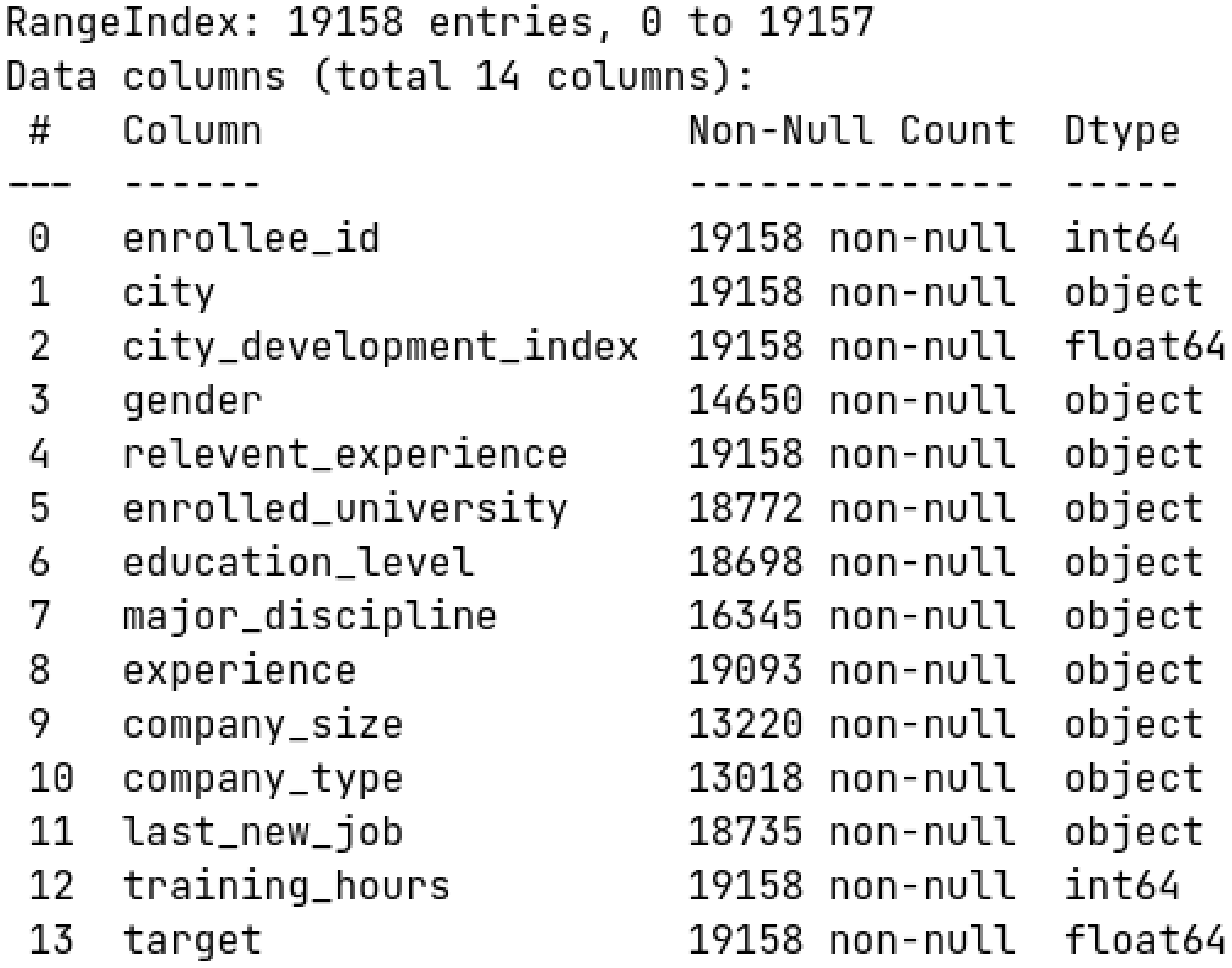
Problem Statement

- Identify candidates intending to work for the company after training.
- Analyze the primary factors influencing employee decisions regarding staying or leaving their current jobs.

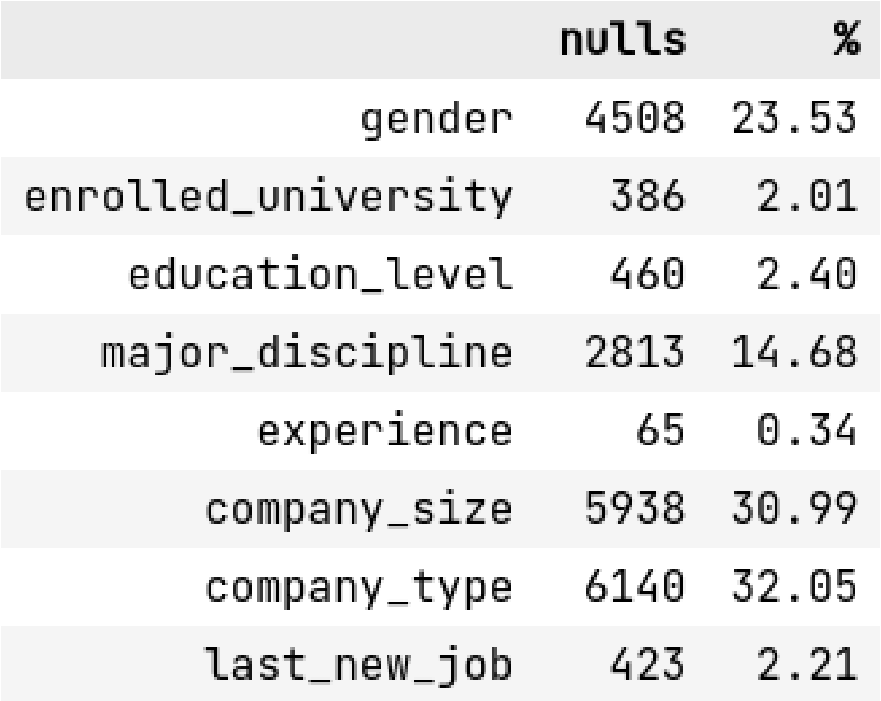
Data



	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new_job	training_hours	target
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	NaN	NaN	1	36	1.0
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>4	47	0.0
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN	NaN	never	83	0.0
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree	<1	NaN	Pvt Ltd	never	52	1.0
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99	Funded Startup	4	8	0.0



```
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   enrollee_id             19158 non-null  int64
 1   city                    19158 non-null  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevent_experience     19158 non-null  object
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  object
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  object
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int64
 13  target                  19158 non-null  float64
```

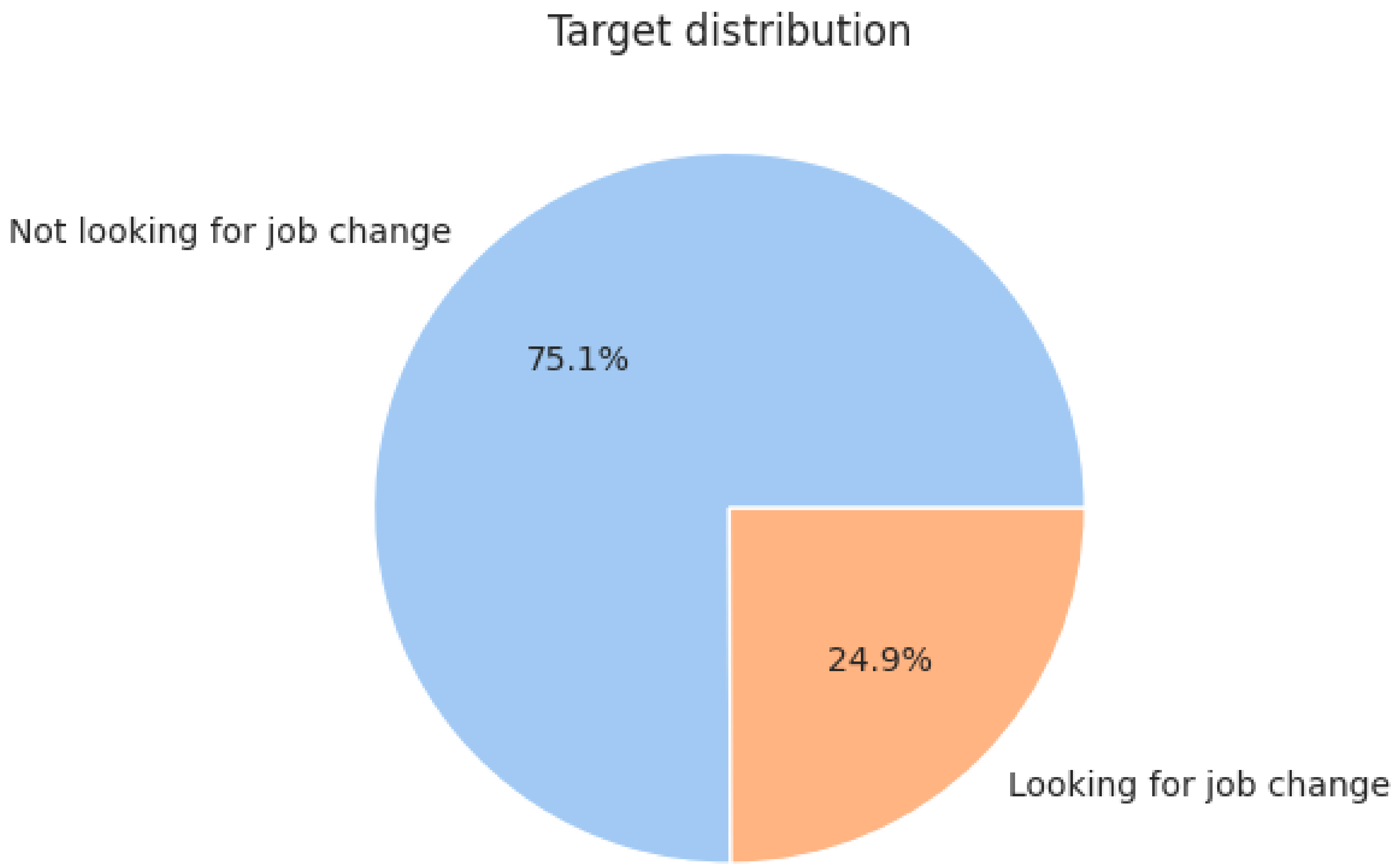


	nulls	%
gender	4508	23.53
enrolled_university	386	2.01
education_level	460	2.40
major_discipline	2813	14.68
experience	65	0.34
company_size	5938	30.99
company_type	6140	32.05
last_new_job	423	2.21

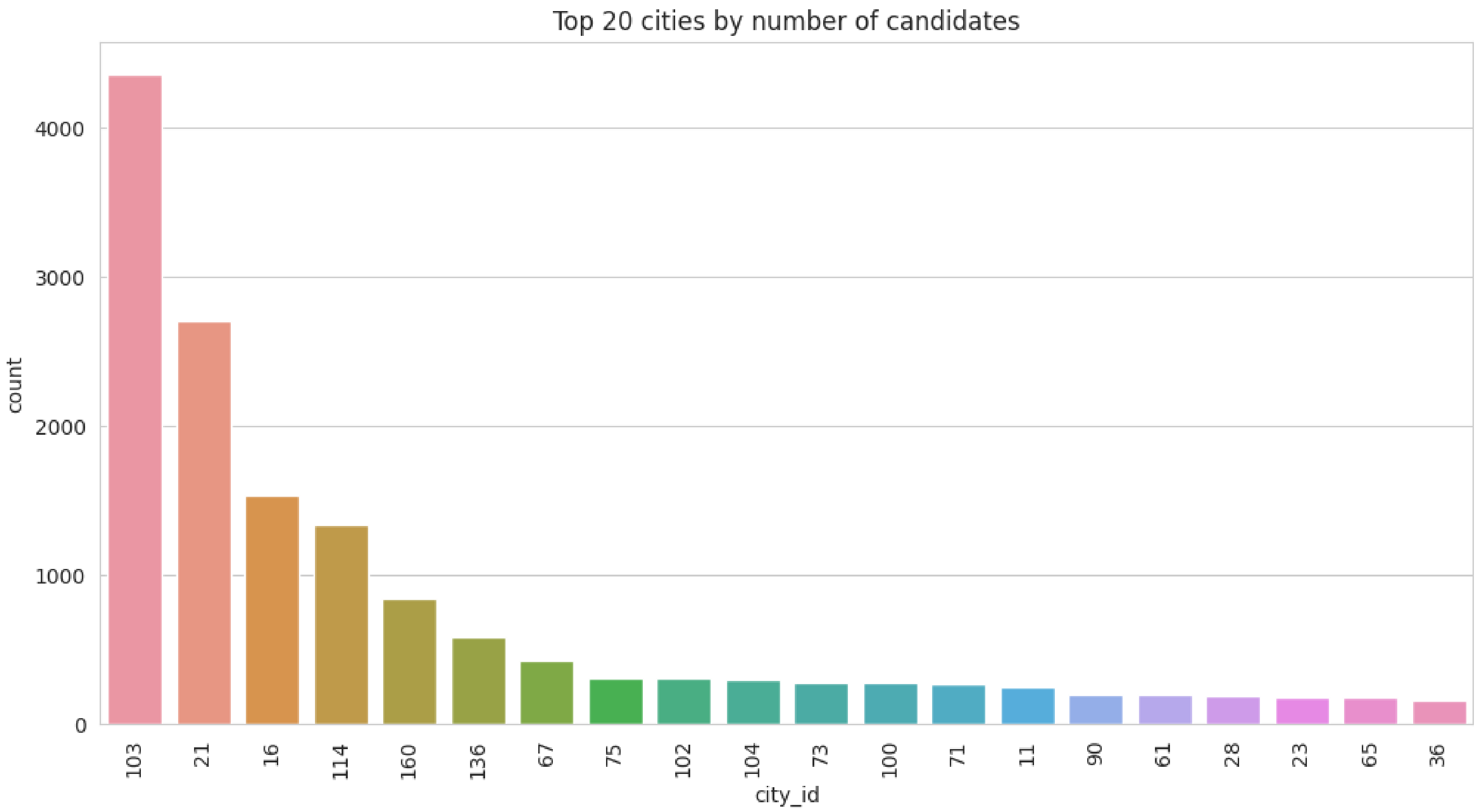
Part 1

EDA Findings

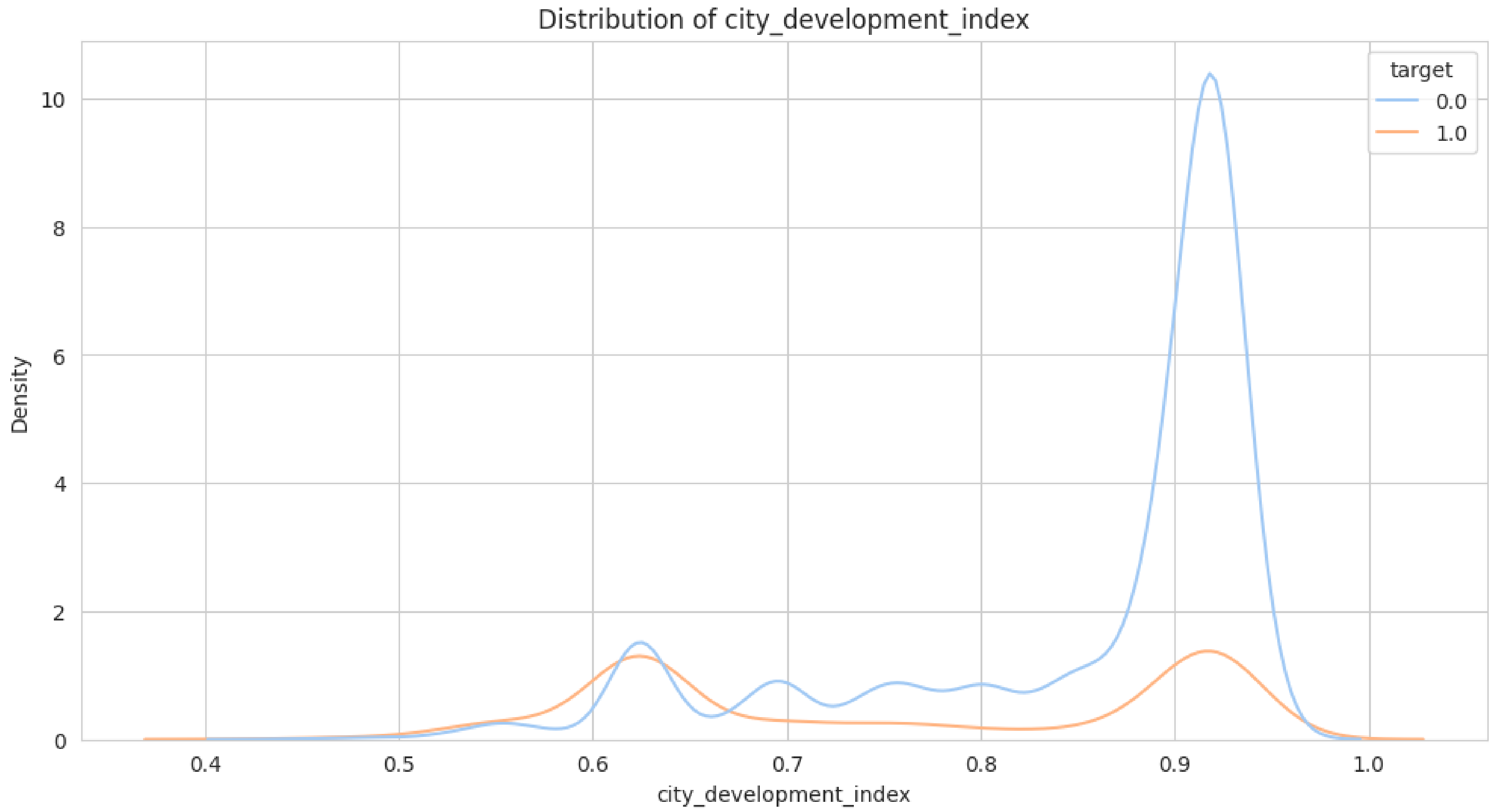
Target Distribution: Unbalanced Data



City Data

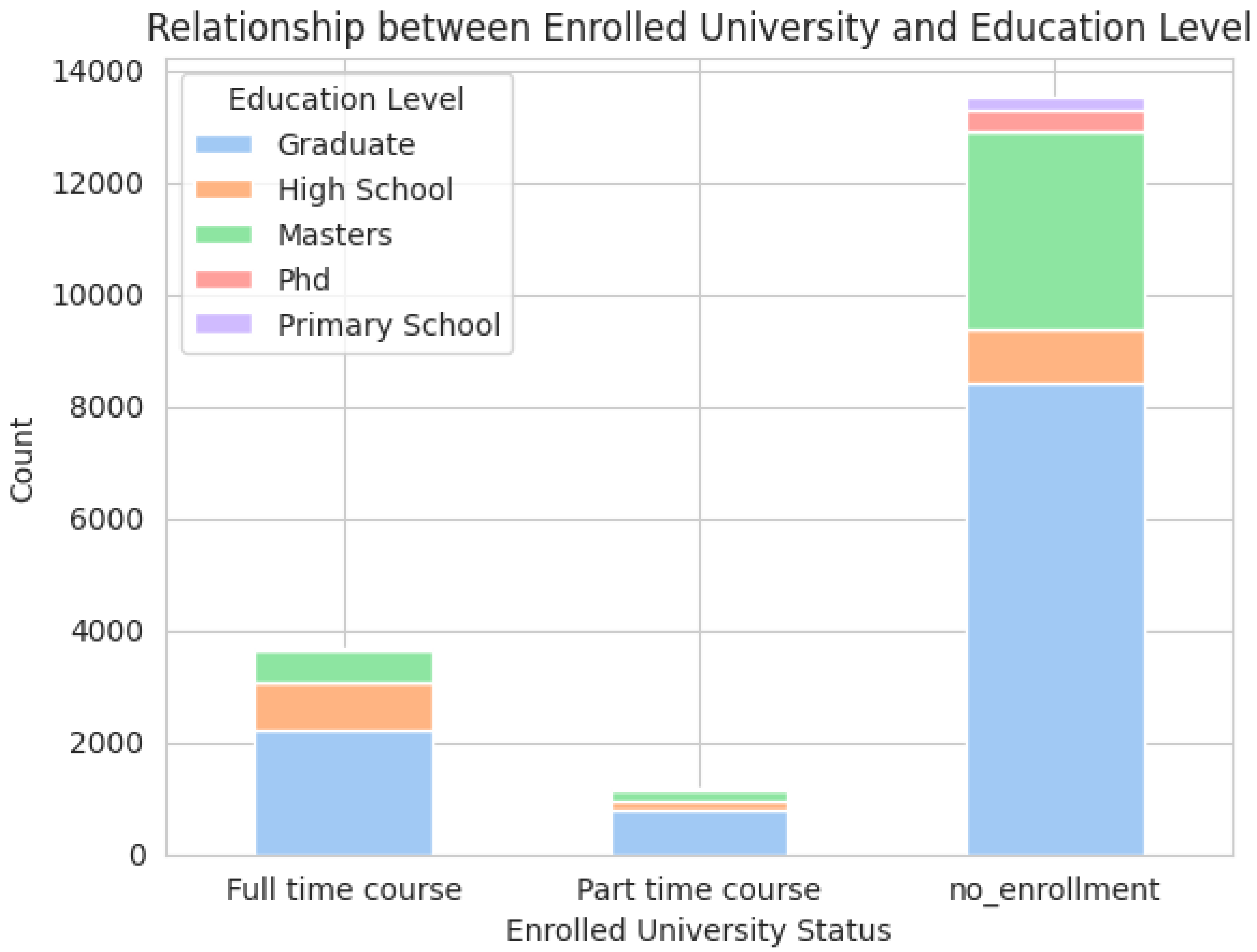


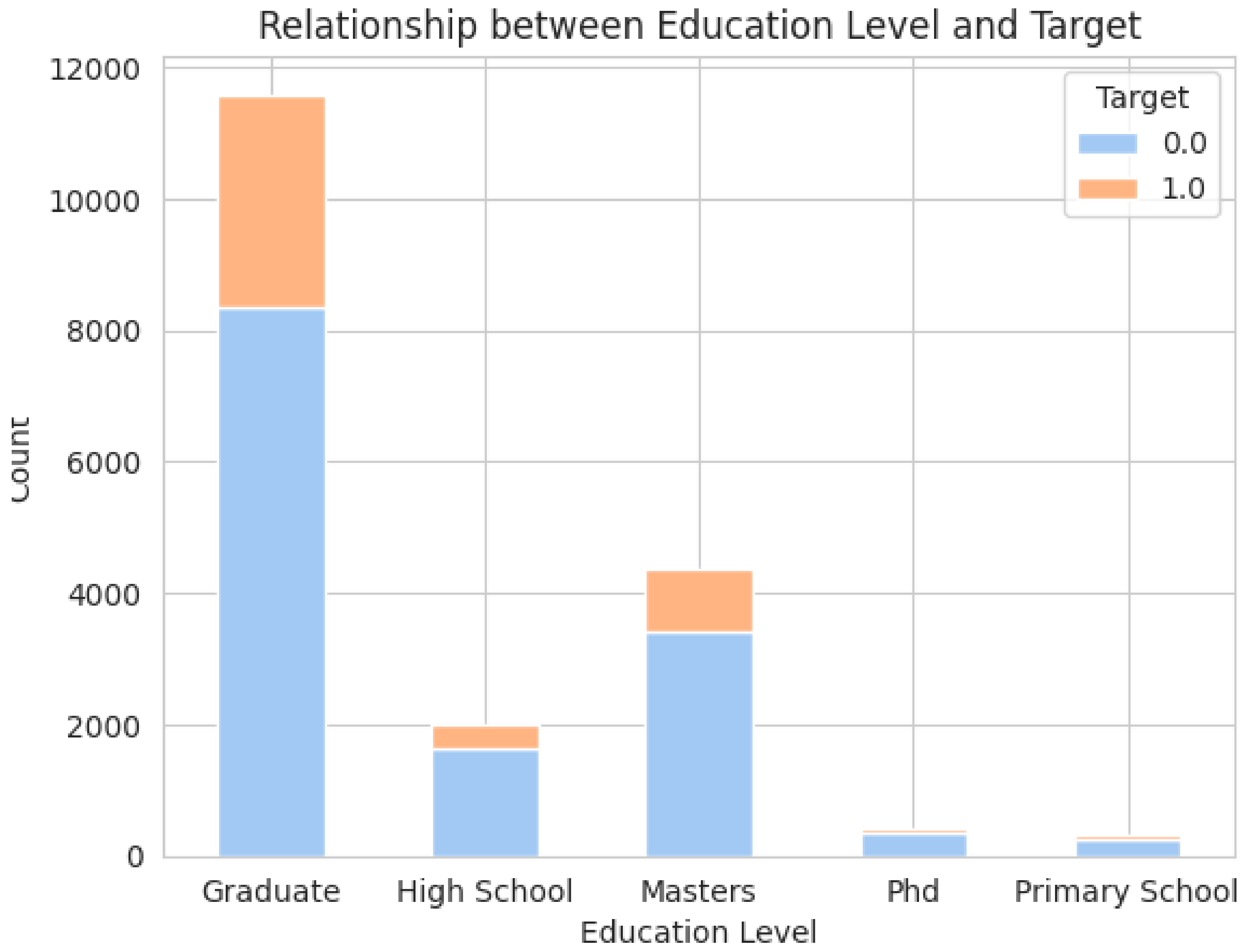
- Most populated cities: 103, 21, 16



- Most candidates are from cities with higher development indices.
- Candidates from cities with higher development indices are less likely to seek a job change.

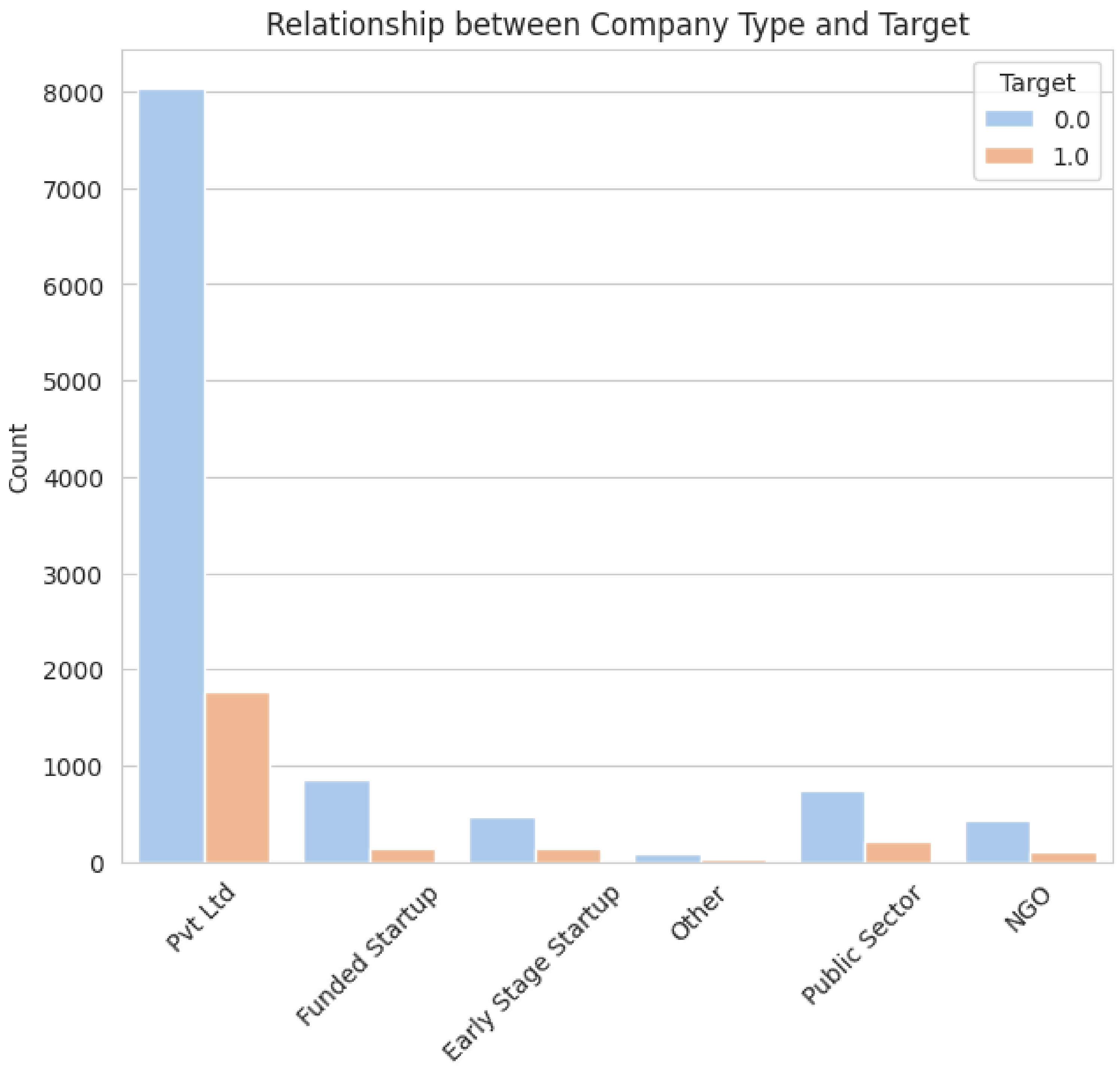
Education data

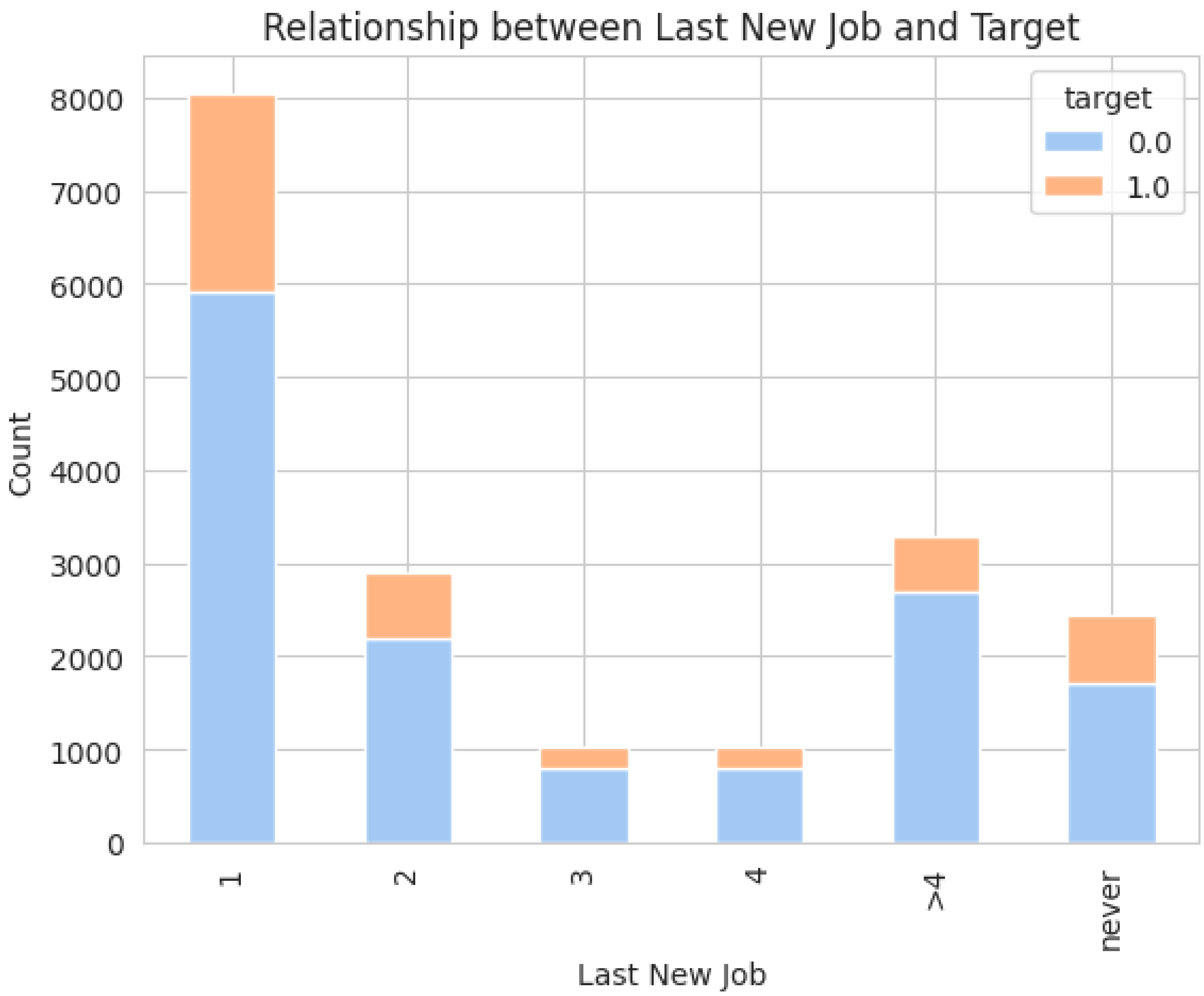




- Around 73% of respondents are not enrolled in any university, and most candidates have a Bachelor's degree.

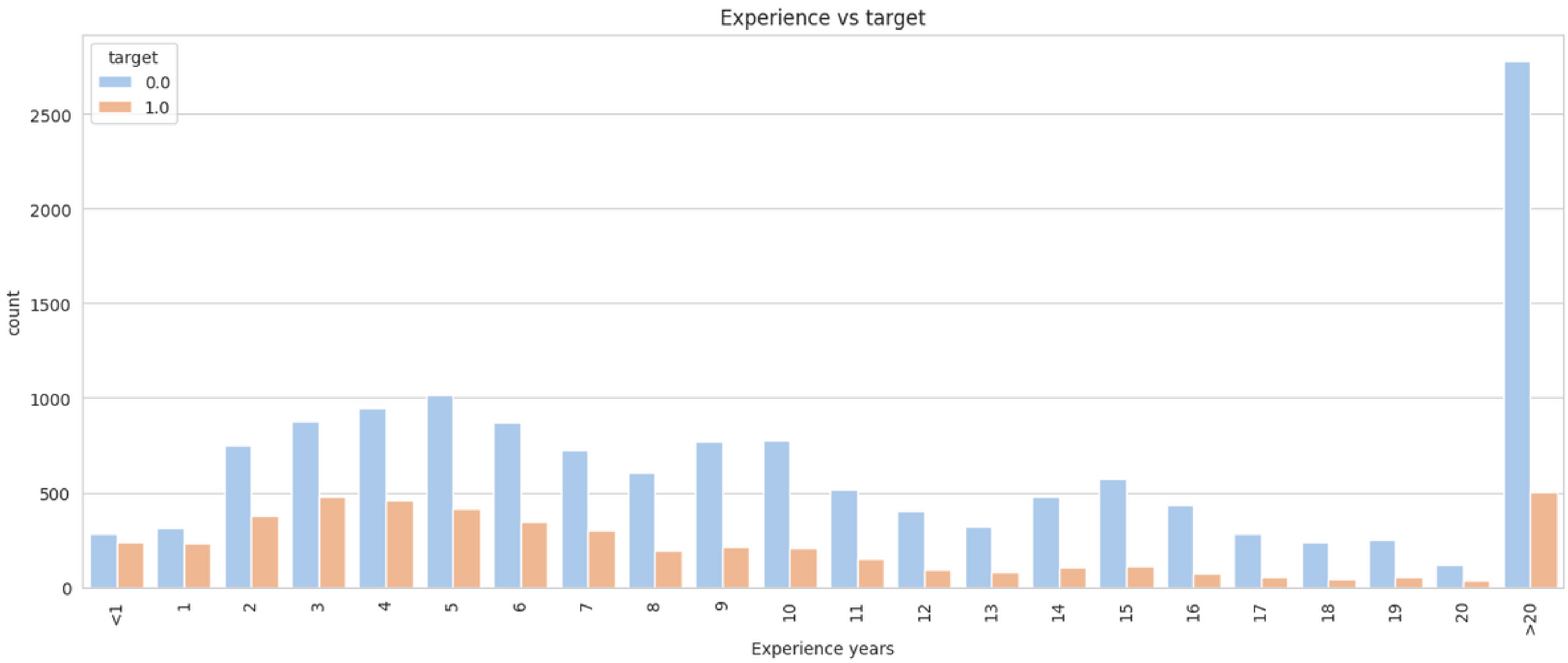
- Graduates are a bit more likely to look for a job change, unlike people with no formal education or PhD.



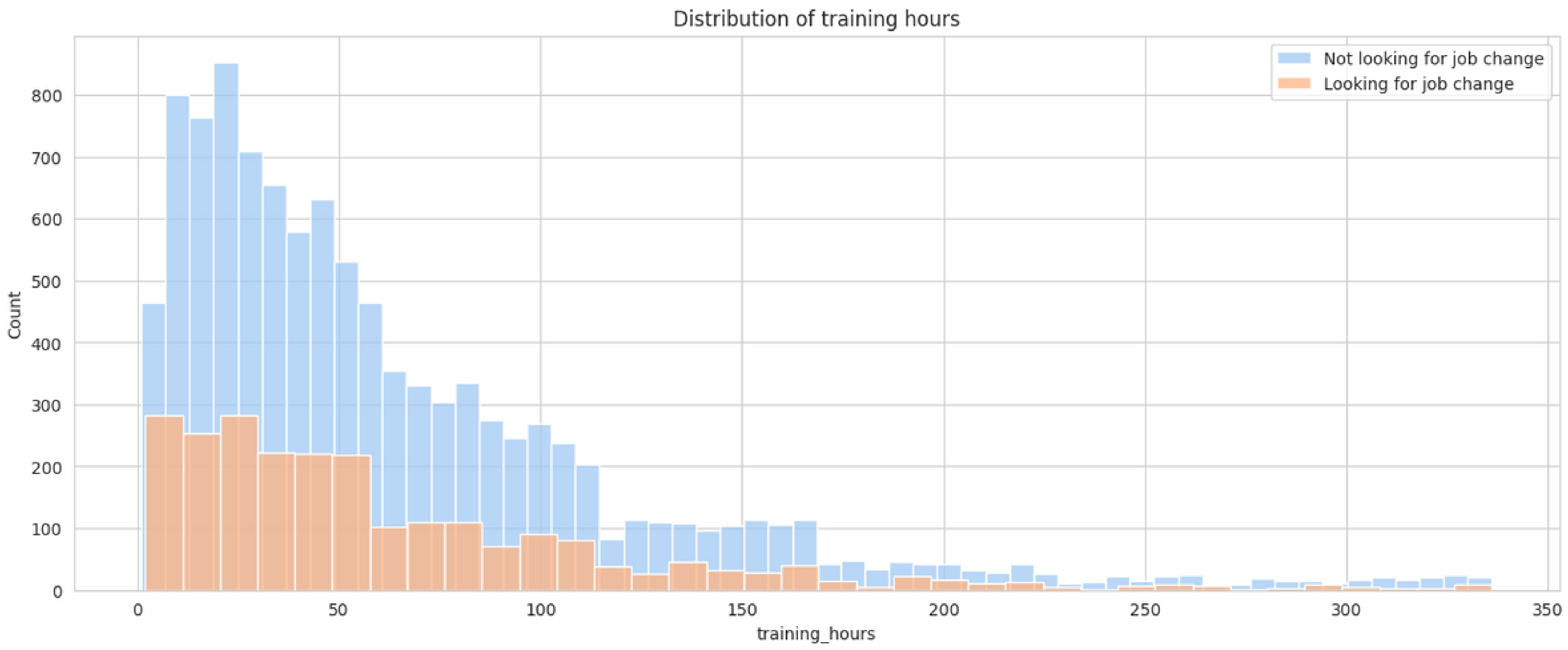


- Most respondents taking courses are employed in Pvt Ltd companies.

- The most common gap between the last and current job is about 1 year. Among them, people are less likely to look for a job change.

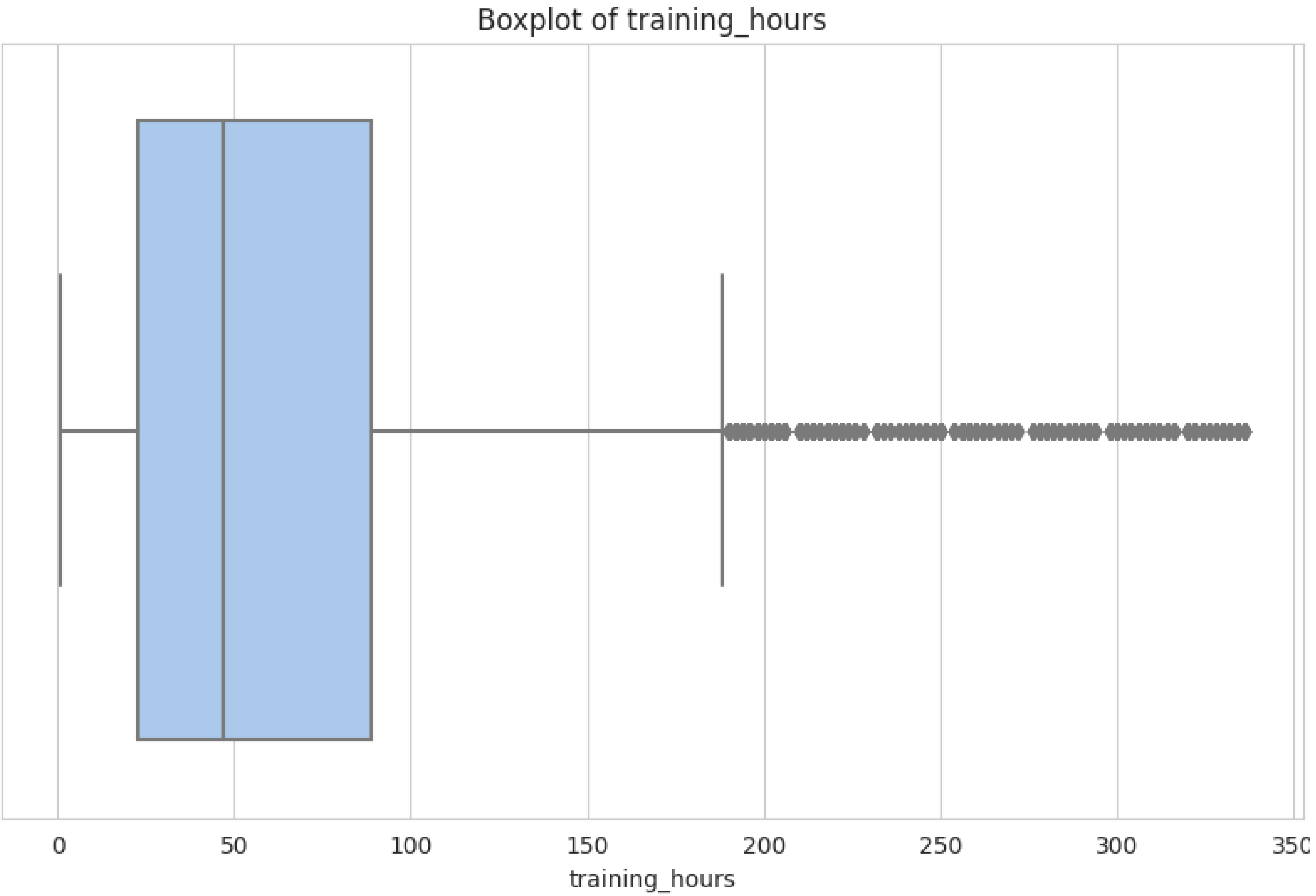


- Most of the people in the dataset have 20+ years of experience.
- Based on distributions (target values for each category), people with 0-4 years of experience are more likely to look for a job change.

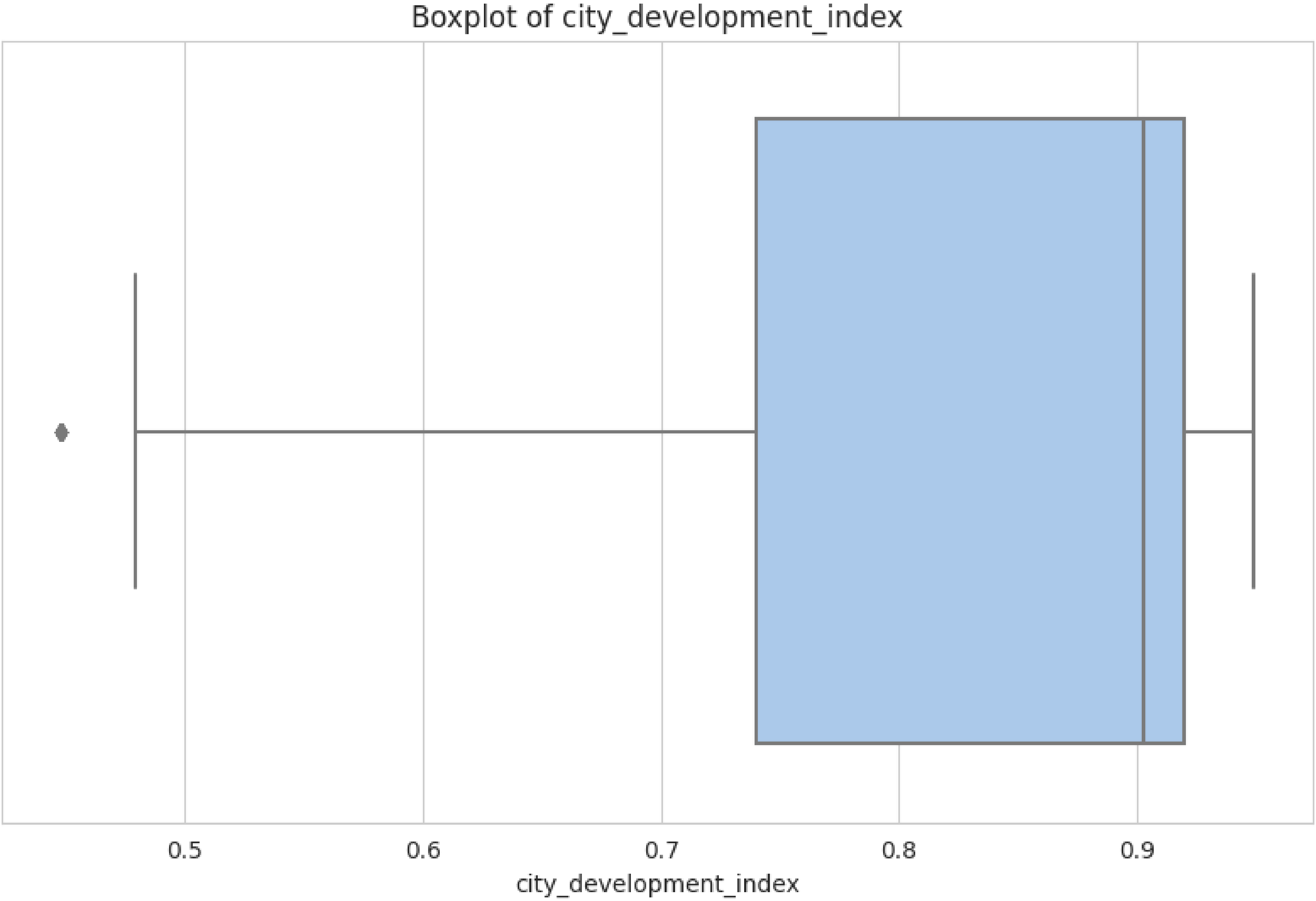


- Most of the people who are looking for a job change have less than 100 hours of training. The peak is at around 20 hours.
- People who are looking for a job change are not necessarily those who are actively learning the most.

Outliers analysis



- 660 outliers: min = 190, max = 336 hours
- Reasonable, did not drop them.



- 17 outliers: city_development_index = 0.448, the lowest value, all corresponding to city_id = 33.
- Reasonable, did not drop them.

Part 2

Data Preprocessing

Handling Missing Values

Drop NaN columns

KNN Imputer

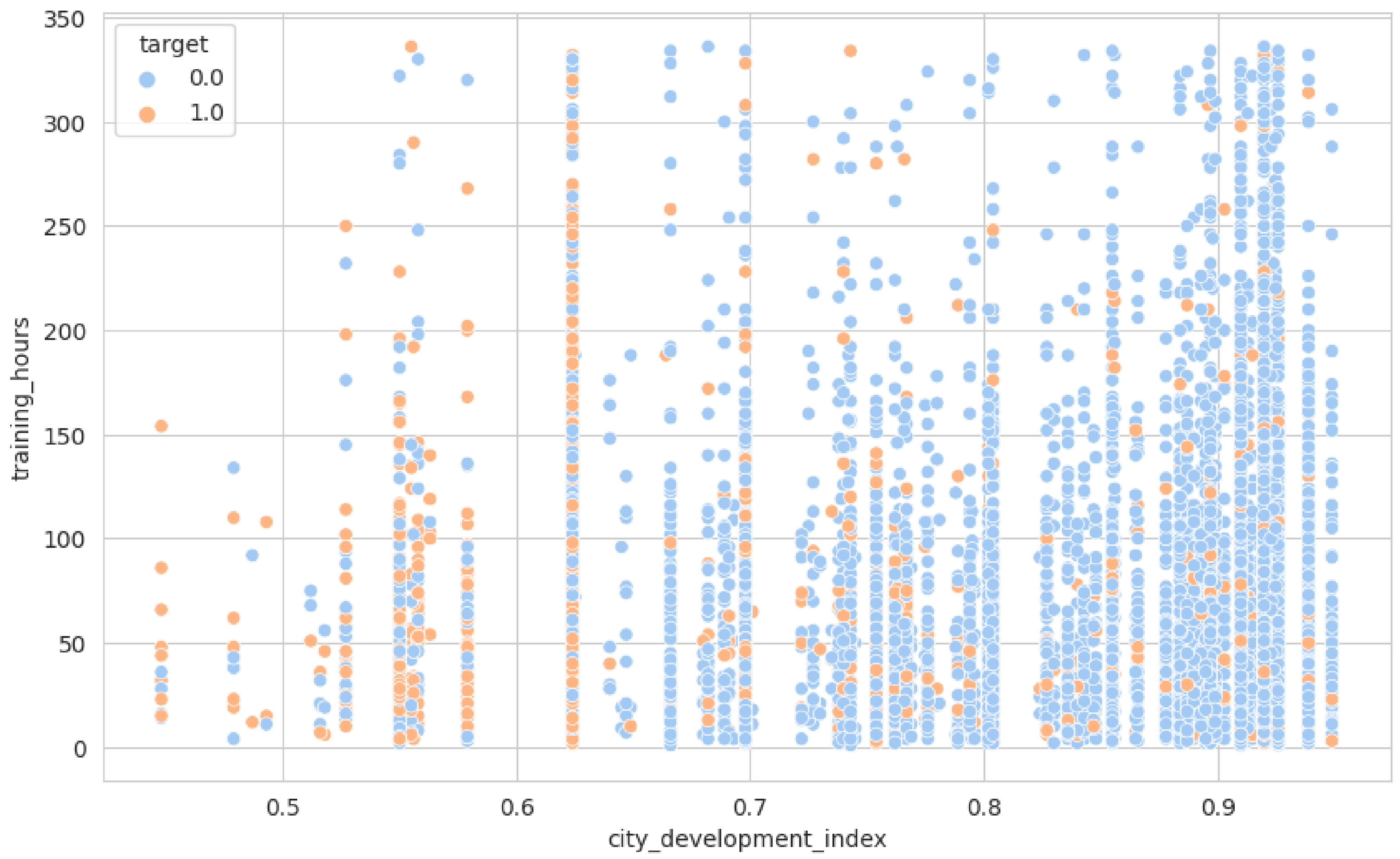
Create new value

Handling Categorical Features

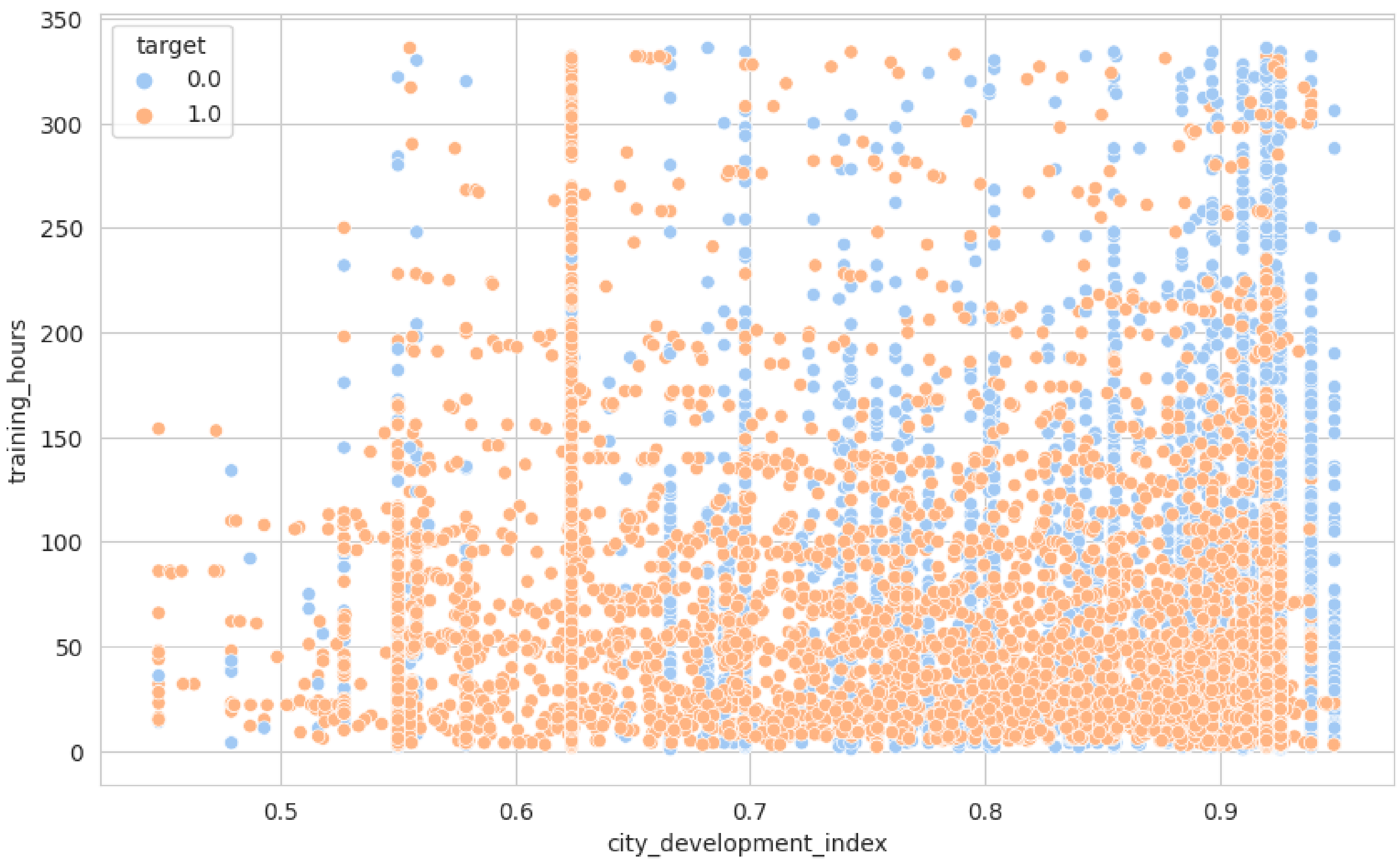
One-hot-encoding

Numeric Conversion

Class Balancing - SMOTE

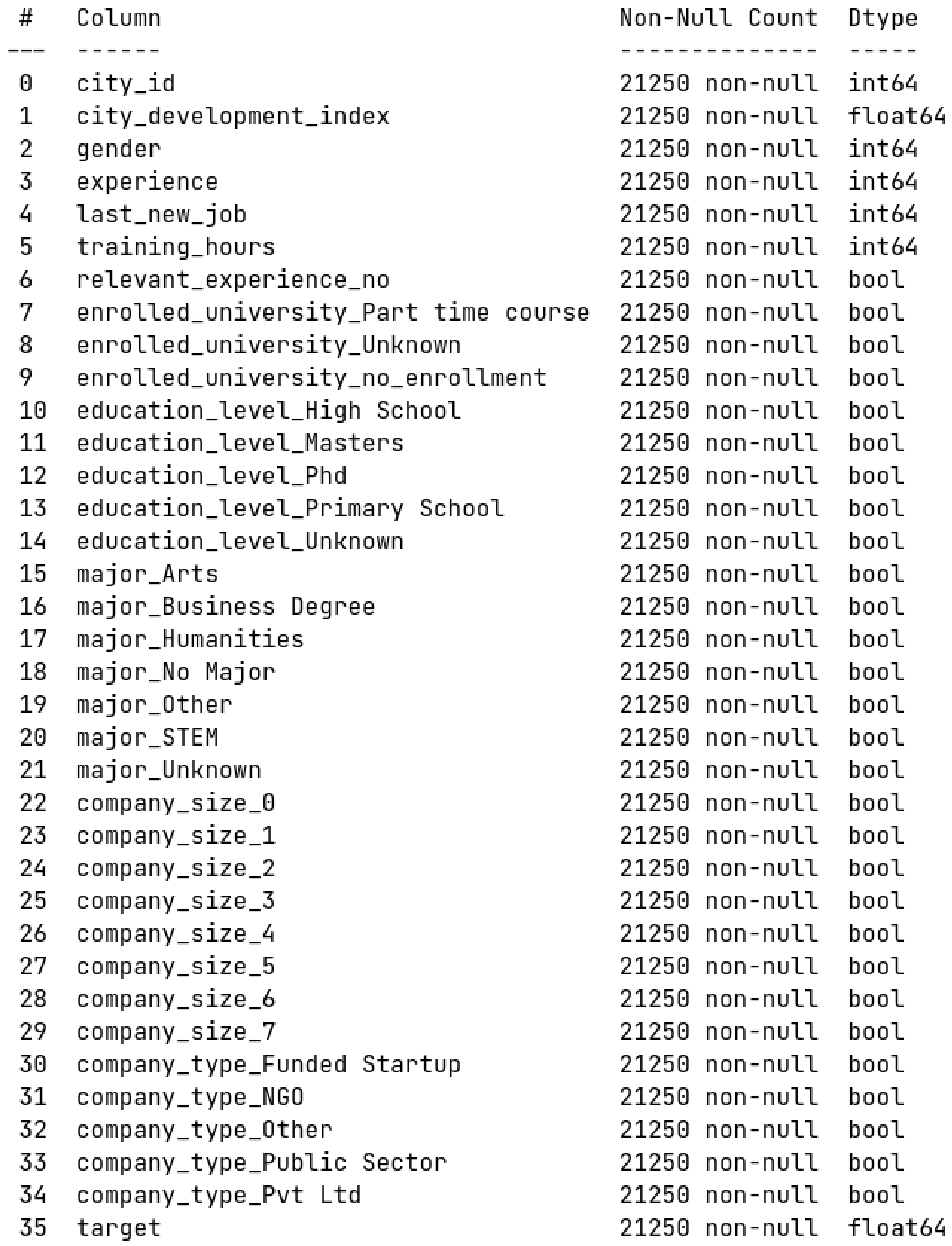


Before SMOTE - 13018 entries



After SMOTE - 21250 entries

Preprocessed Data

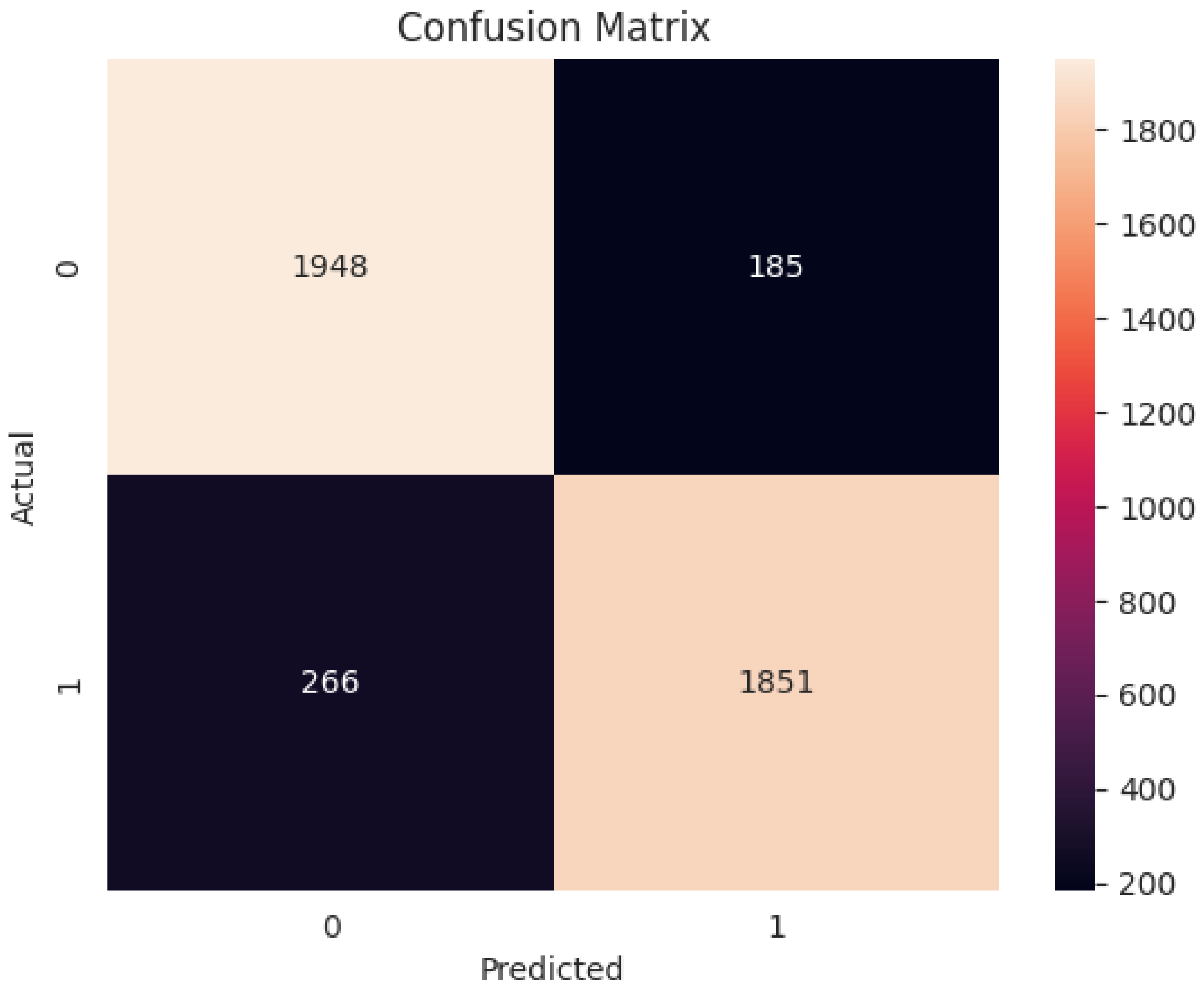


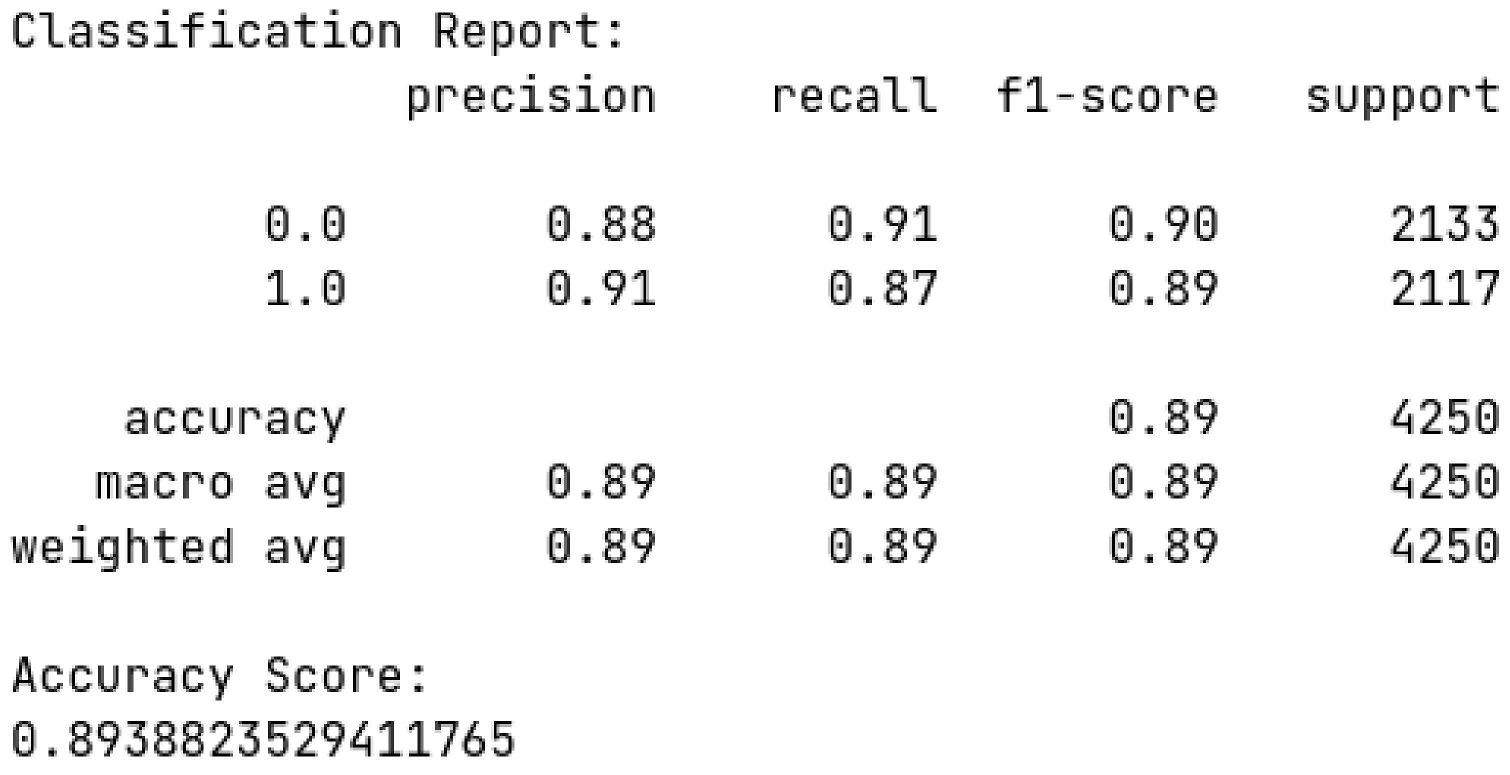
```
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   city_id                             21250 non-null  int64
 1   city_development_index              21250 non-null  float64
 2   gender                              21250 non-null  int64
 3   experience                          21250 non-null  int64
 4   last_new_job                        21250 non-null  int64
 5   training_hours                      21250 non-null  int64
 6   relevant_experience_no              21250 non-null  bool
 7   enrolled_university_Part time course  21250 non-null  bool
 8   enrolled_university_Unknown         21250 non-null  bool
 9   enrolled_university_no_enrollment   21250 non-null  bool
 10  education_level_High School         21250 non-null  bool
 11  education_level_Masters             21250 non-null  bool
 12  education_level_Phd                 21250 non-null  bool
 13  education_level_Primary School      21250 non-null  bool
 14  education_level_Unknown             21250 non-null  bool
 15  major_Arts                          21250 non-null  bool
 16  major_Business Degree               21250 non-null  bool
 17  major_Humanities                    21250 non-null  bool
 18  major_No Major                      21250 non-null  bool
 19  major_Other                         21250 non-null  bool
 20  major_STEM                          21250 non-null  bool
 21  major_Unknown                       21250 non-null  bool
 22  company_size_0                      21250 non-null  bool
 23  company_size_1                      21250 non-null  bool
 24  company_size_2                      21250 non-null  bool
 25  company_size_3                      21250 non-null  bool
 26  company_size_4                      21250 non-null  bool
 27  company_size_5                      21250 non-null  bool
 28  company_size_6                      21250 non-null  bool
 29  company_size_7                      21250 non-null  bool
 30  company_type_Funded Startup         21250 non-null  bool
 31  company_type_NGO                    21250 non-null  bool
 32  company_type_Other                  21250 non-null  bool
 33  company_type_Public Sector          21250 non-null  bool
 34  company_type_Pvt Ltd                21250 non-null  bool
 35  target                              21250 non-null  float64
```

Part 3

Modelling

Random Forest

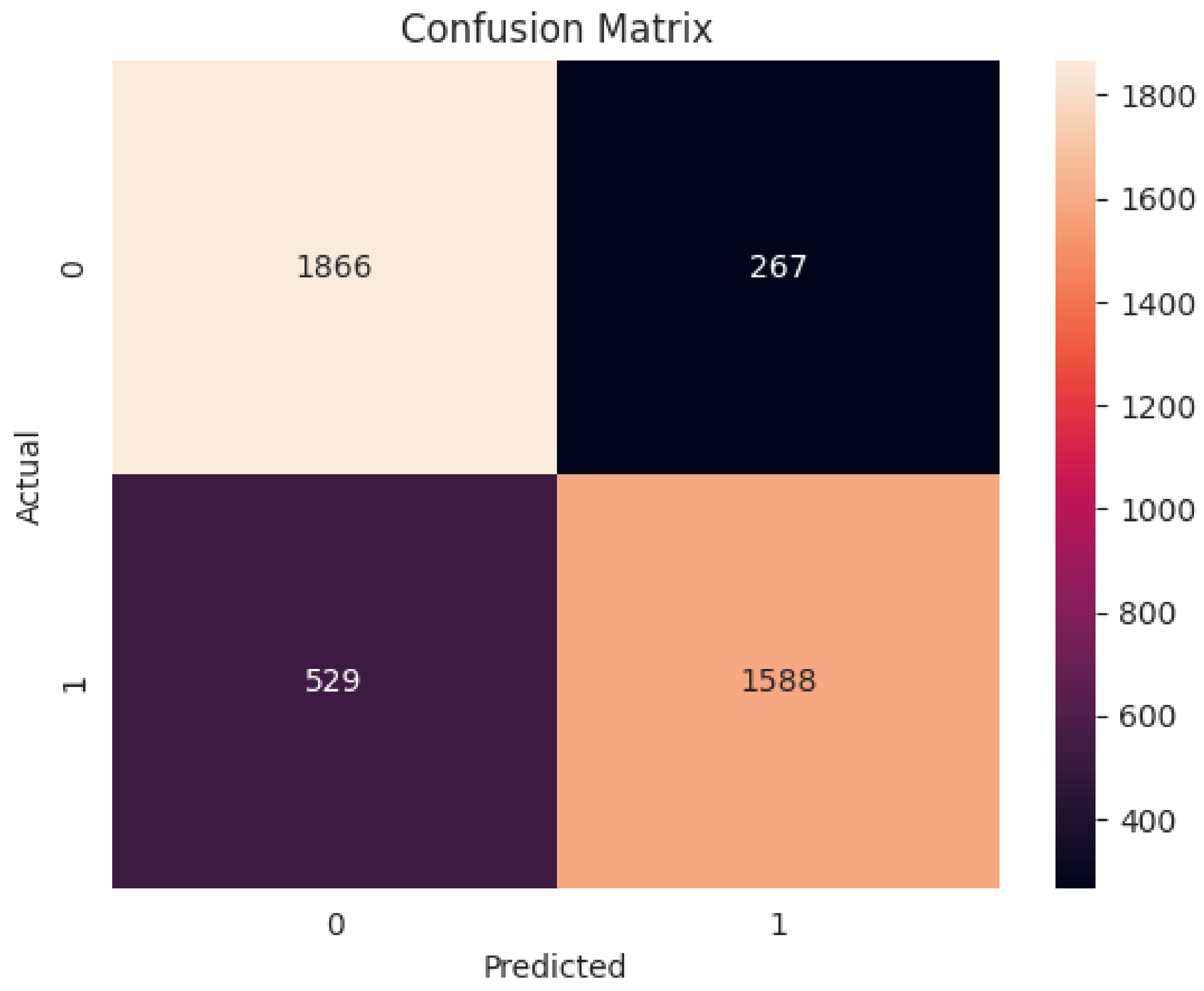


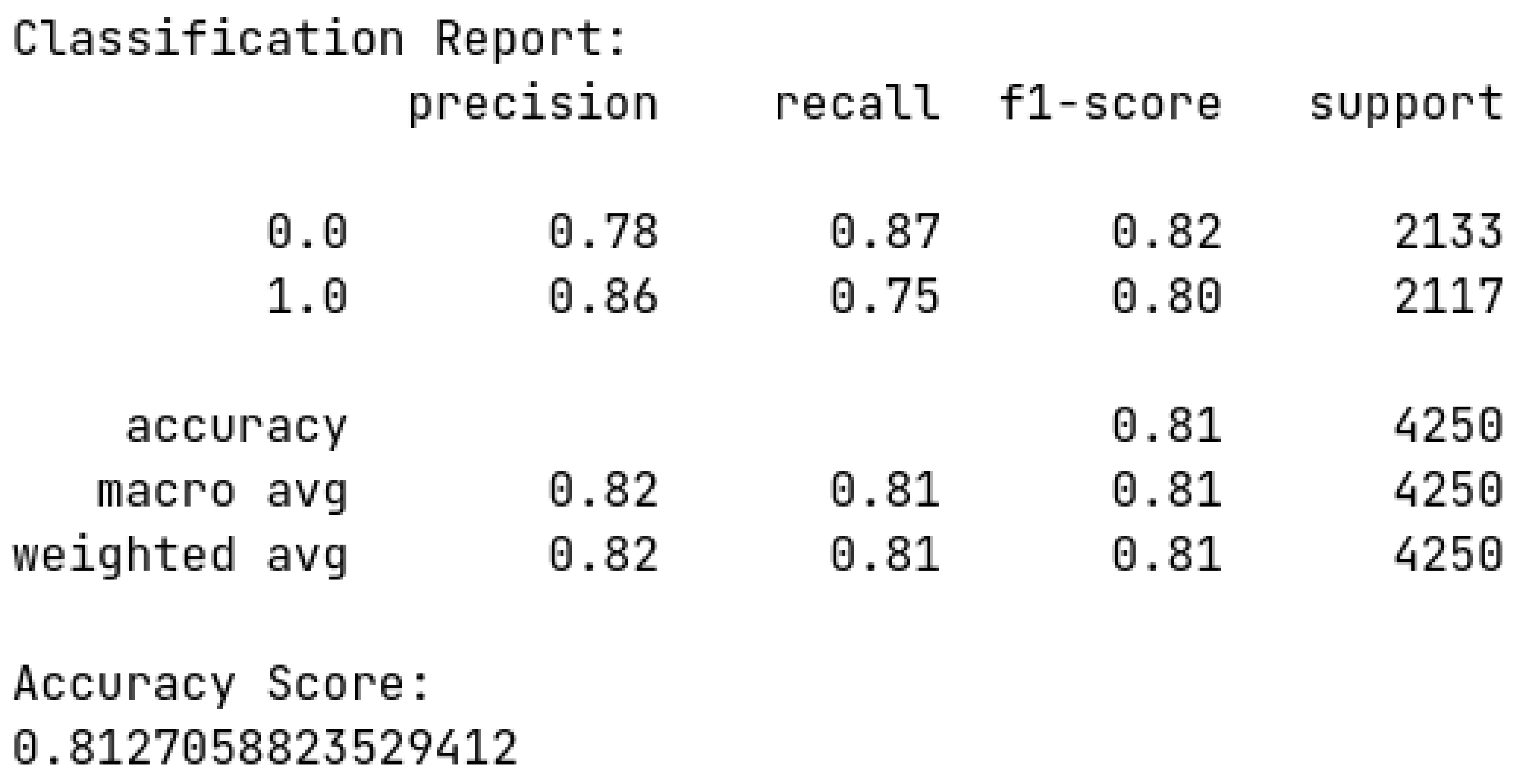


```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.91      0.90      2133
         1.0       0.91      0.87      0.89      2117

    accuracy                           0.89      4250
   macro avg       0.89      0.89      0.89      4250
weighted avg       0.89      0.89      0.89      4250

Accuracy Score:
0.8938823529411765
```

Feature Selection*

The Point-Biserial Correlation Selector: ['city_development_index', 'city_id', 'experience']



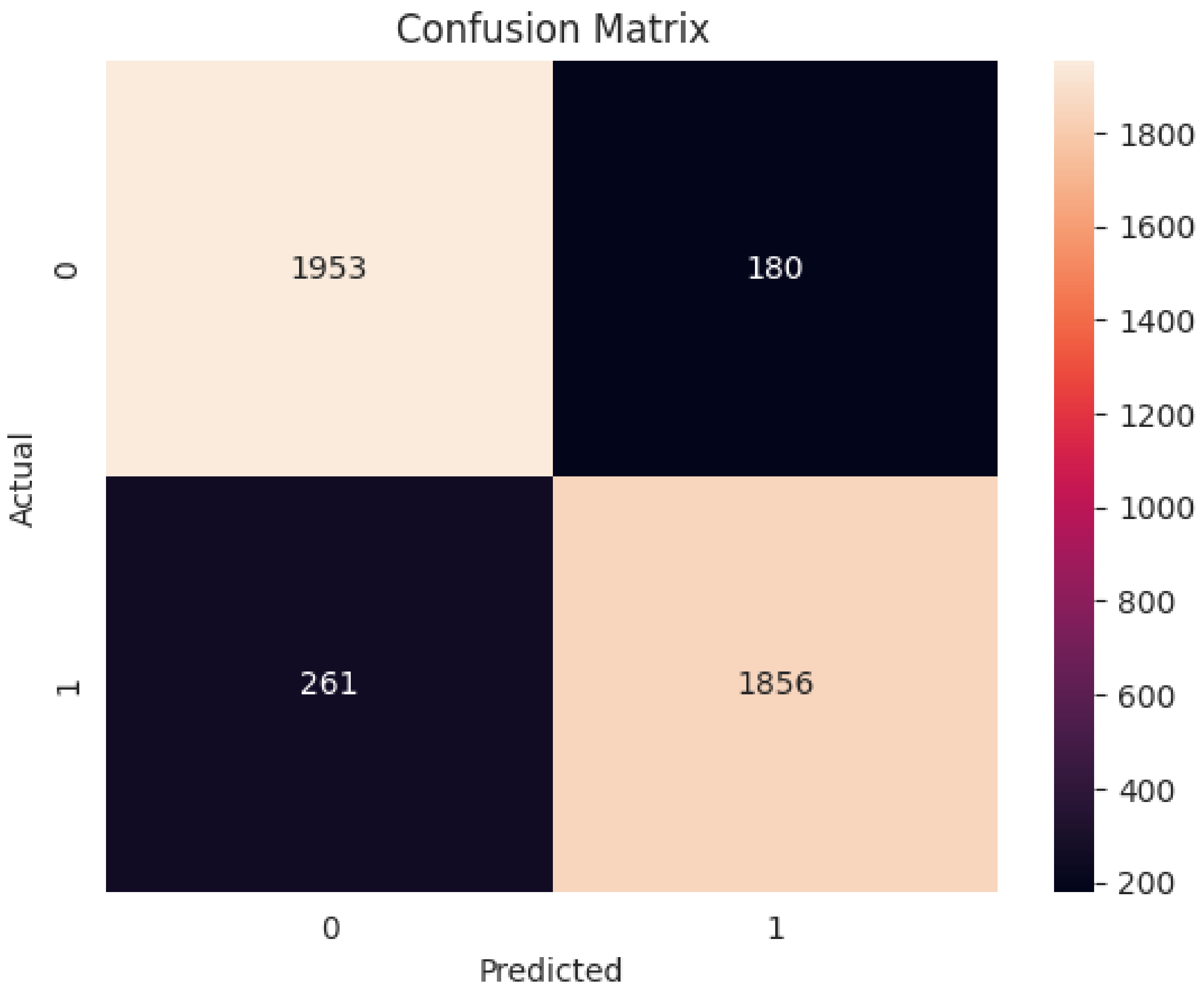


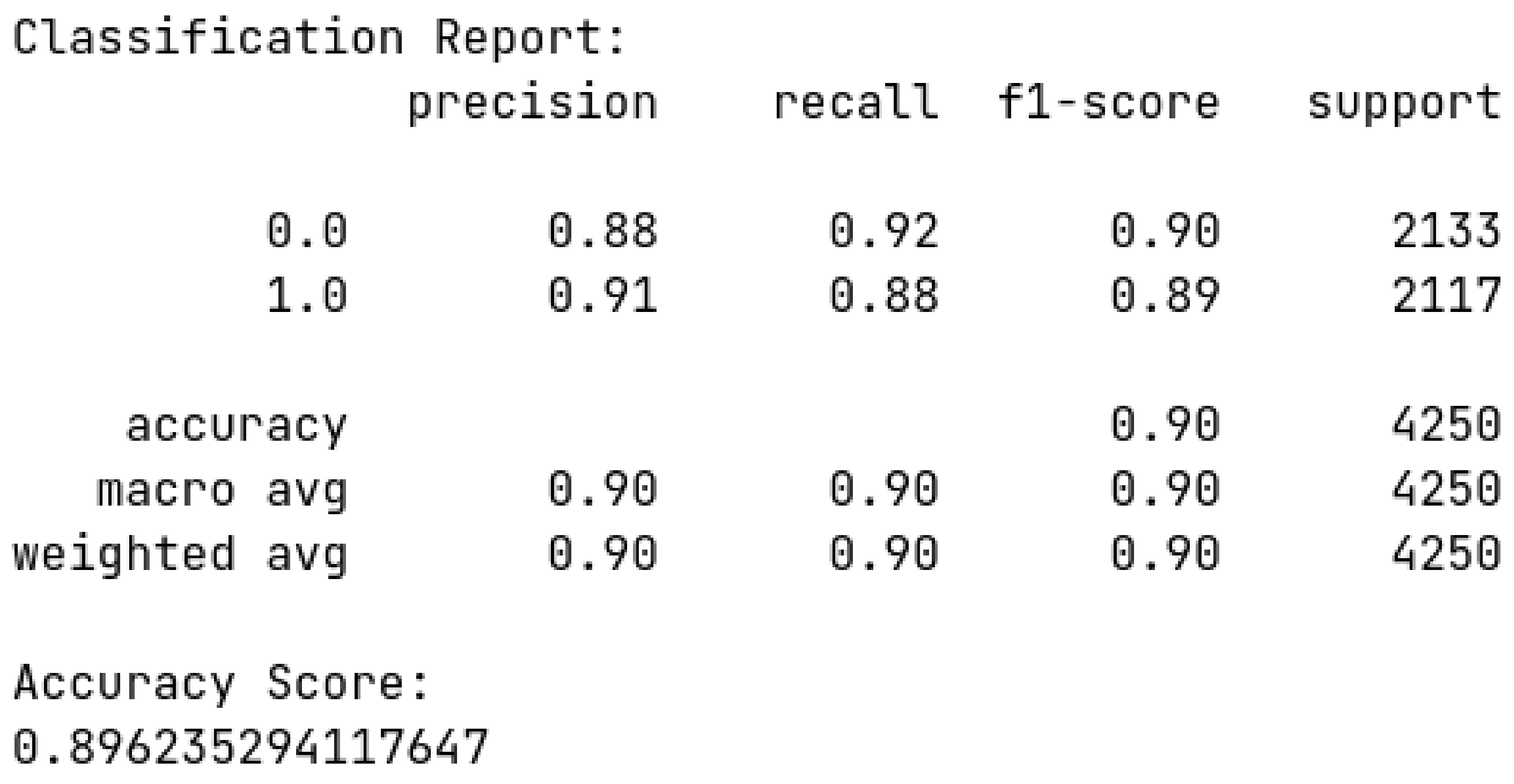
```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.78      0.87      0.82      2133
         1.0       0.86      0.75      0.80      2117

    accuracy                           0.81      4250
   macro avg       0.82      0.81      0.81      4250
weighted avg       0.82      0.81      0.81      4250

Accuracy Score:
0.8127058823529412
```

*Did not apply to the final model

Hyperparameter Tunning

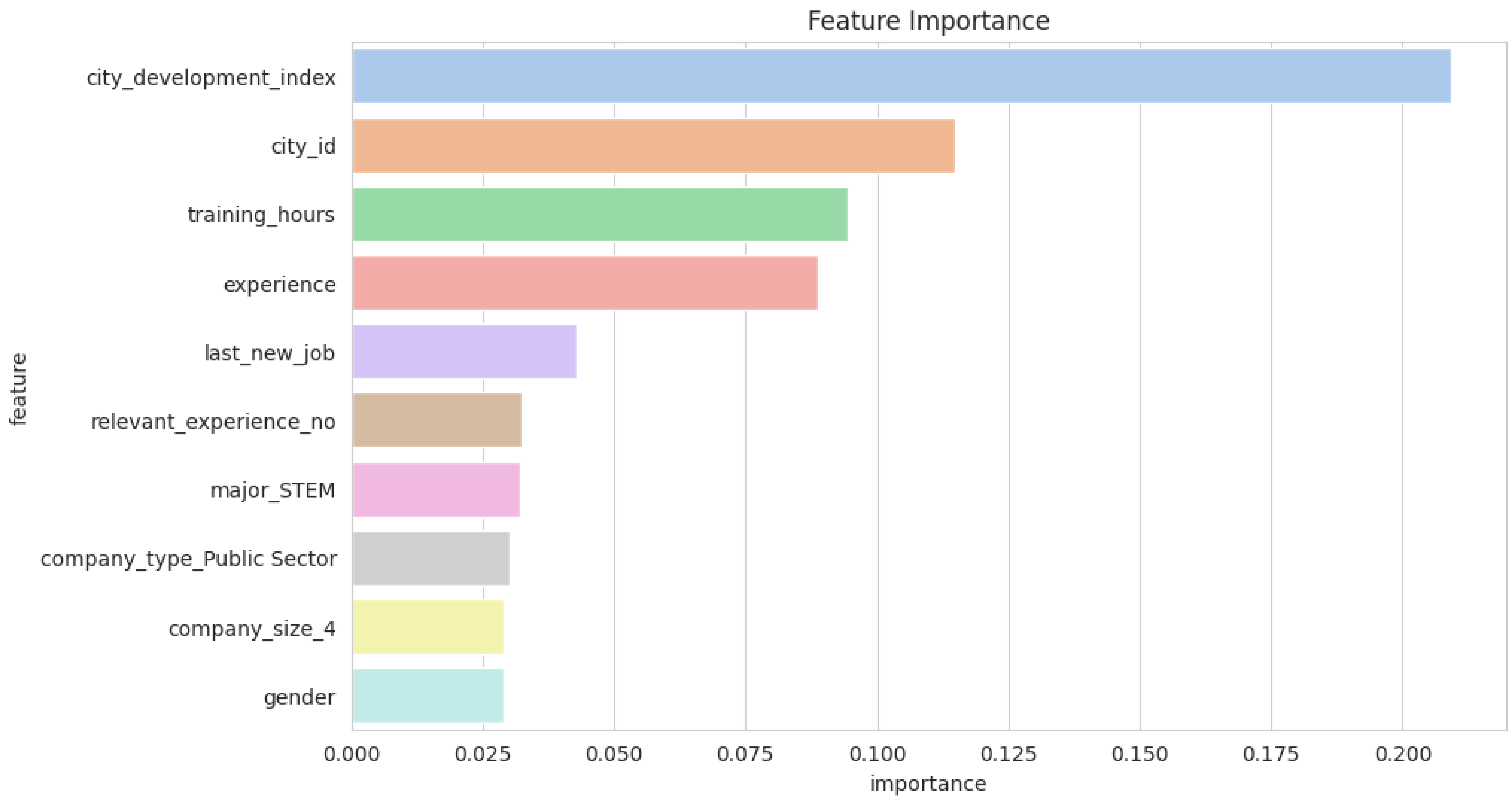
HalvingGridSearchCV Parameters: {'bootstrap': False, 'criterion': 'gini', 'max_depth': 60, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 400}



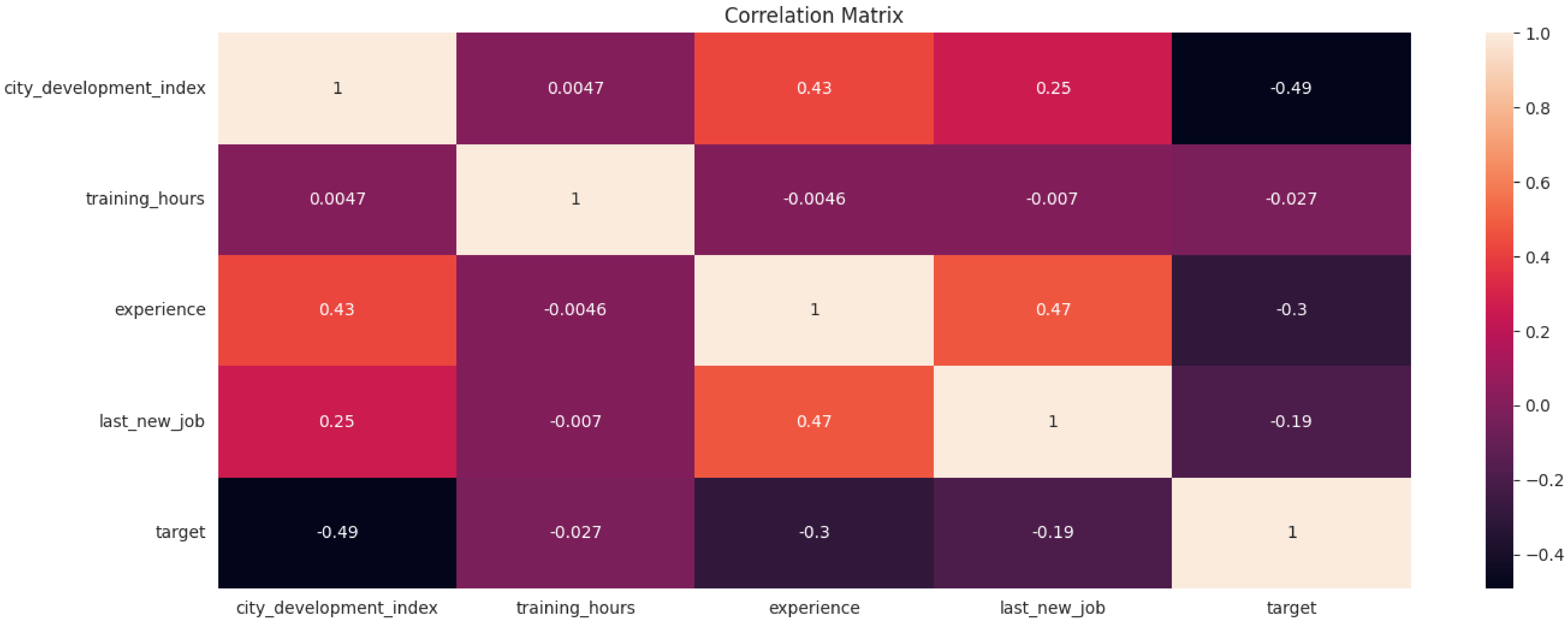


```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.92      0.90      2133
         1.0       0.91      0.88      0.89      2117

    accuracy                           0.90      4250
   macro avg       0.90      0.90      0.90      4250
weighted avg       0.90      0.90      0.90      4250

Accuracy Score:
0.896235294117647
```

Feature Importance



Correlation Matrix



	city_development_index	training_hours	experience	last_new_job	target
city_development_index	1	0.0047	0.43	0.25	-0.49
training_hours	0.0047	1	-0.0046	-0.007	-0.027
experience	0.43	-0.0046	1	0.47	-0.3
last_new_job	0.25	-0.007	0.47	1	-0.19
target	-0.49	-0.027	-0.3	-0.19	1

Conclusion

The primary factors influencing employees' decision to change their jobs are

- city development index
- years of experience
- completed training hours
- difference in years between previous and current job.