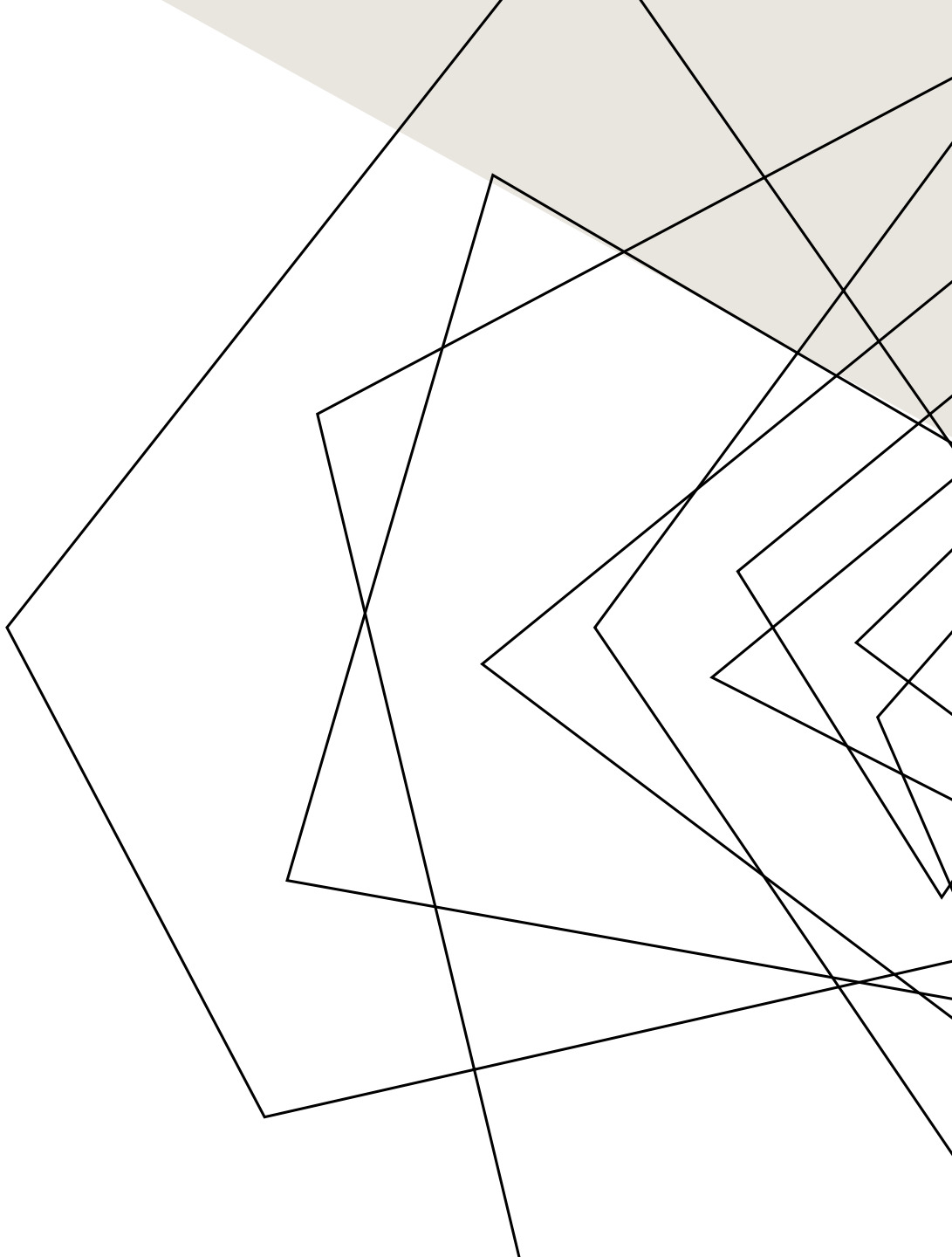
DATA SCIENCE

SALARY TRENDS

MACHINE LEARNING MODELS

AGENDA

- Dataset Introduction
- Univariate Analysis
- Bivariate Analysis
- Machine Learning Models

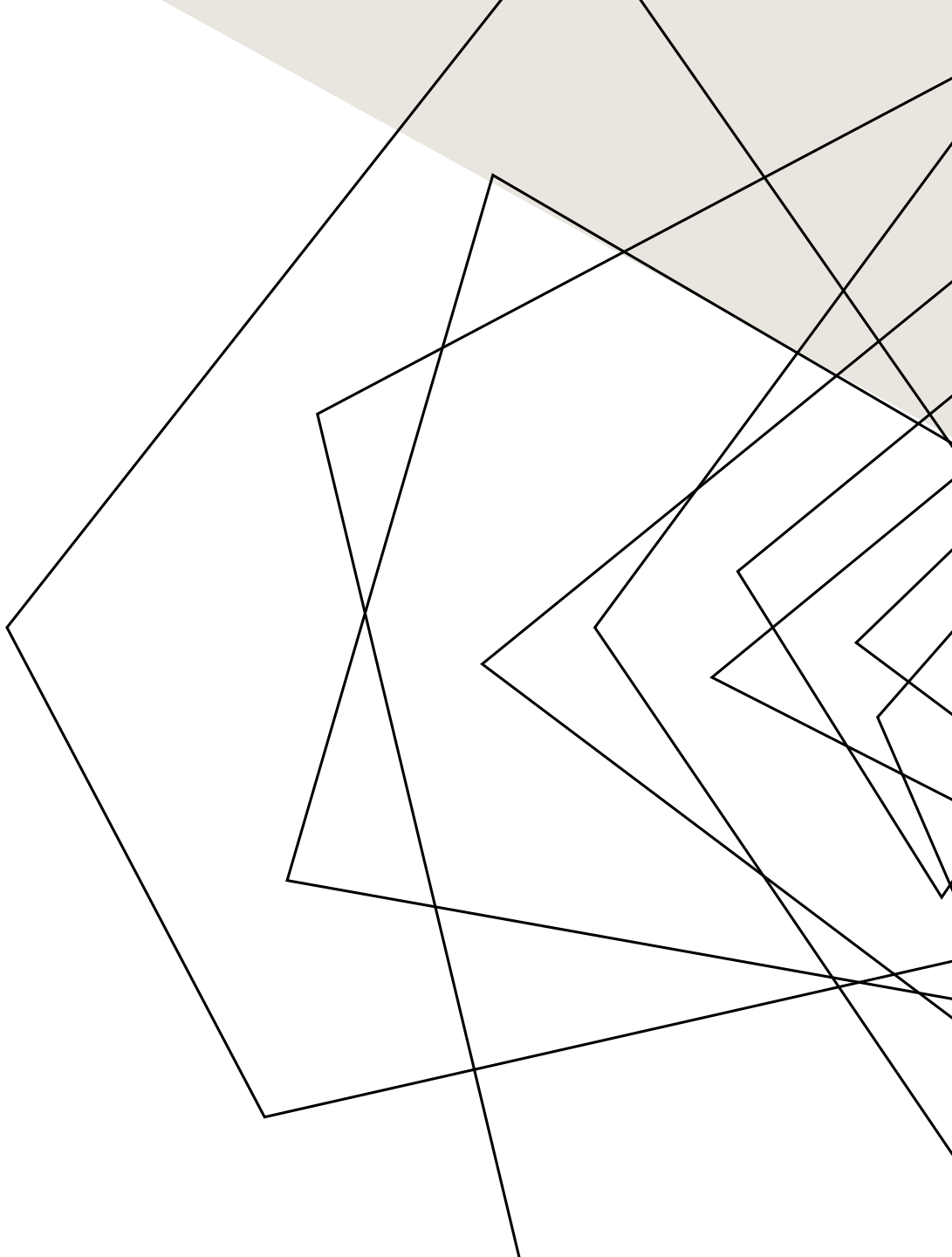


DATASET INTRODUCTION

- Total Observations: 3755
- Missing Values: NO
- Duplicates Values: 1171
- Numerical Variables: 2
  - salary
  - salary_in_usd
- Categorical Variables: 9
  - work_year (2020-2023) (4)
  - experience_level (4)
  - employment_type (4)
  - salary_currency (20)
  - employee_residence (78)
  - remote_ratio (3)
  - company_location (72)
  - company_size (3)
  - job_category (4)

```
Index: 2584 entries, 0 to 3754
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   work_year               2584 non-null   int64
 1   experience_level        2584 non-null   object
 2   full_time               2584 non-null   int64
 3   salary                  2584 non-null   int64
 4   salary_currency         2584 non-null   object
 5   salary_in_usd           2584 non-null   int64
 6   employee_residence(US)  2584 non-null   int64
 7   remote_ratio            2584 non-null   object
 8   company_location(US)    2584 non-null   int64
 9   company_size            2584 non-null   object
 10  job_category            2584 non-null   object
 11  salary_bins             2584 non-null   category
dtypes: category(1), int64(6), object(5)
memory usage: 244.9+ KB
```

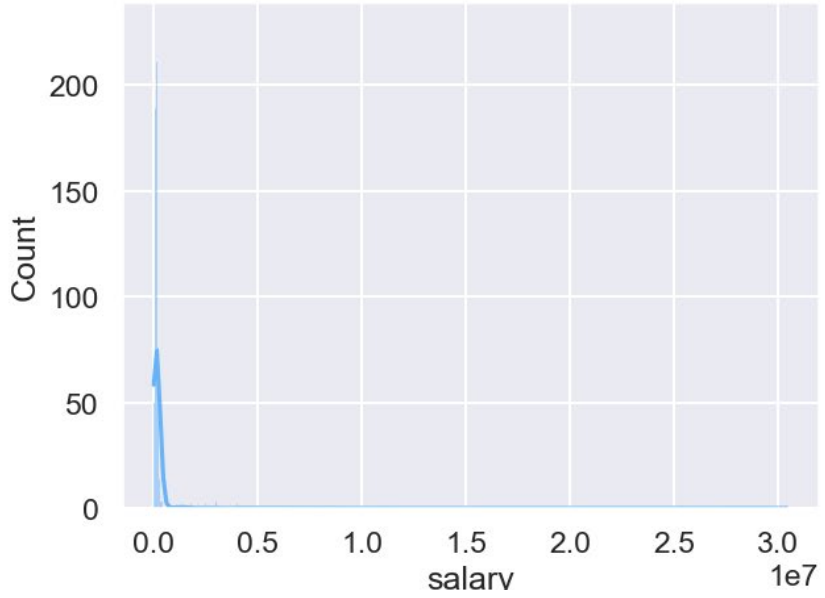
UNIVARIATE ANALYSIS

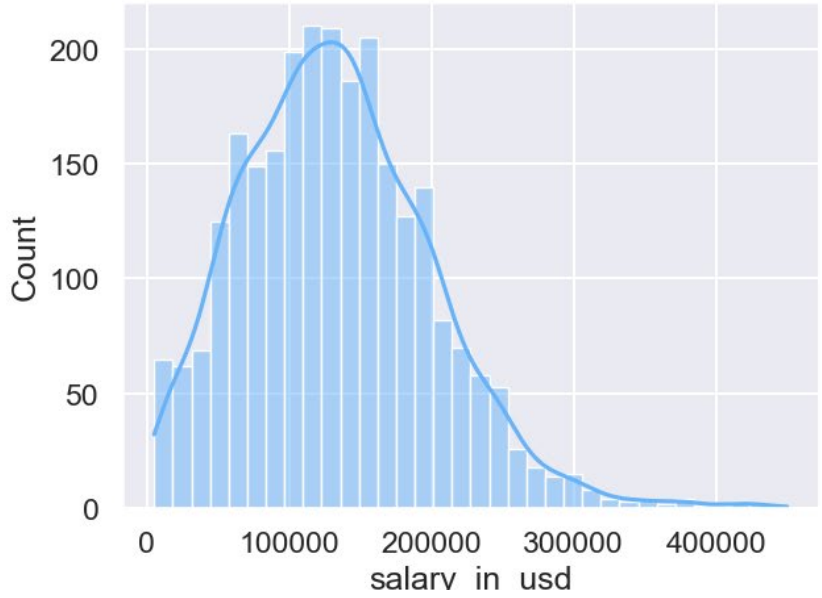


NUMERICAL VARIABLES

salary

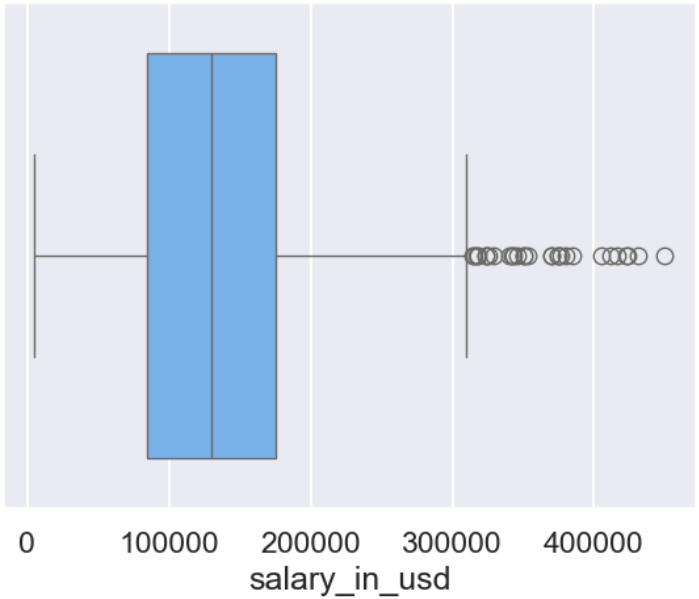
- Skewness: 24.09
- CV: 3.84





salary_in_usd

- Skewness: 0.62
- CV: 0.5
- Min: 5132
- Mean: 130k
- Median: 133k
- Max: 450k



SALARY_IN_USD

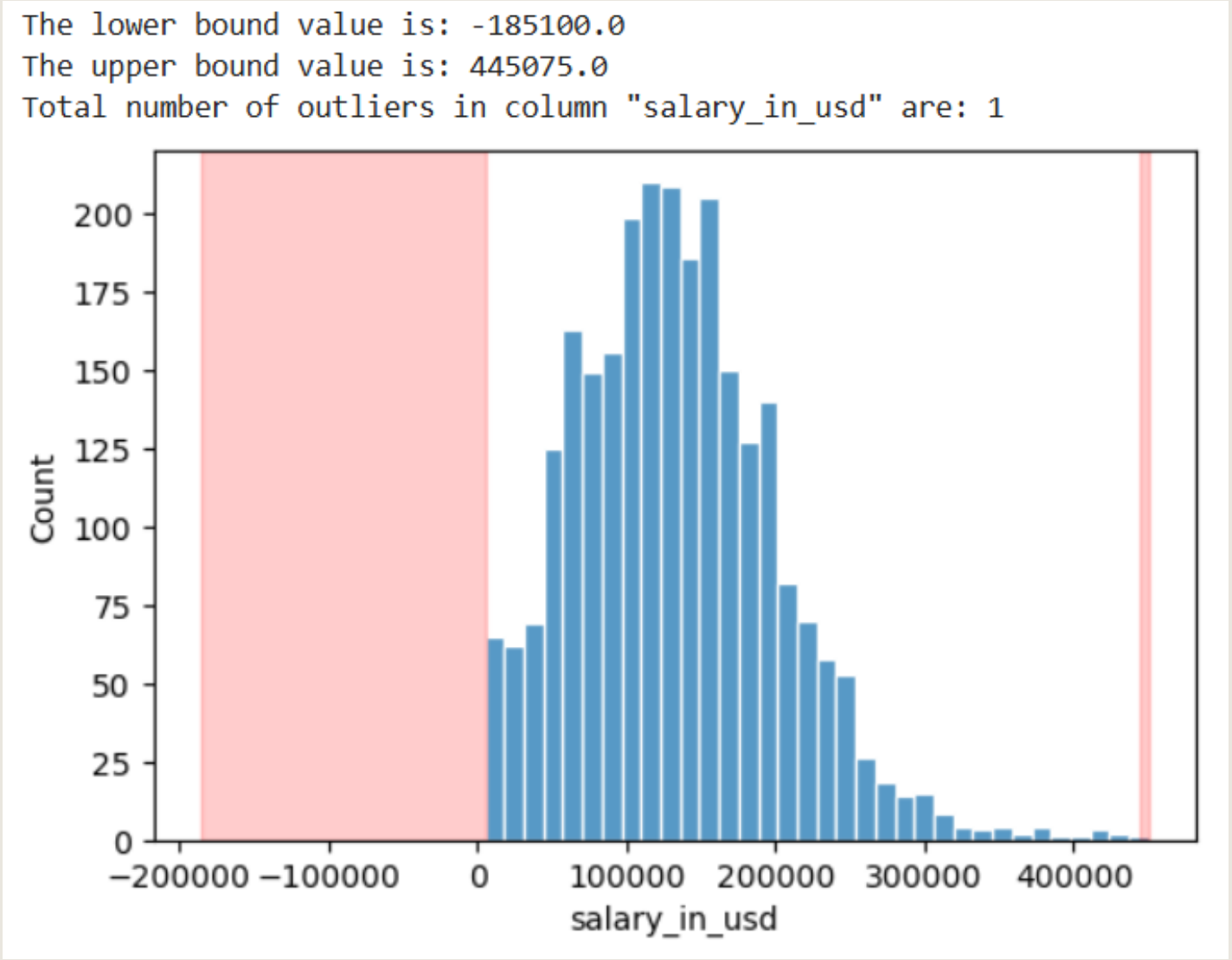
Outliers Detection

Turkey Method

Target Variable

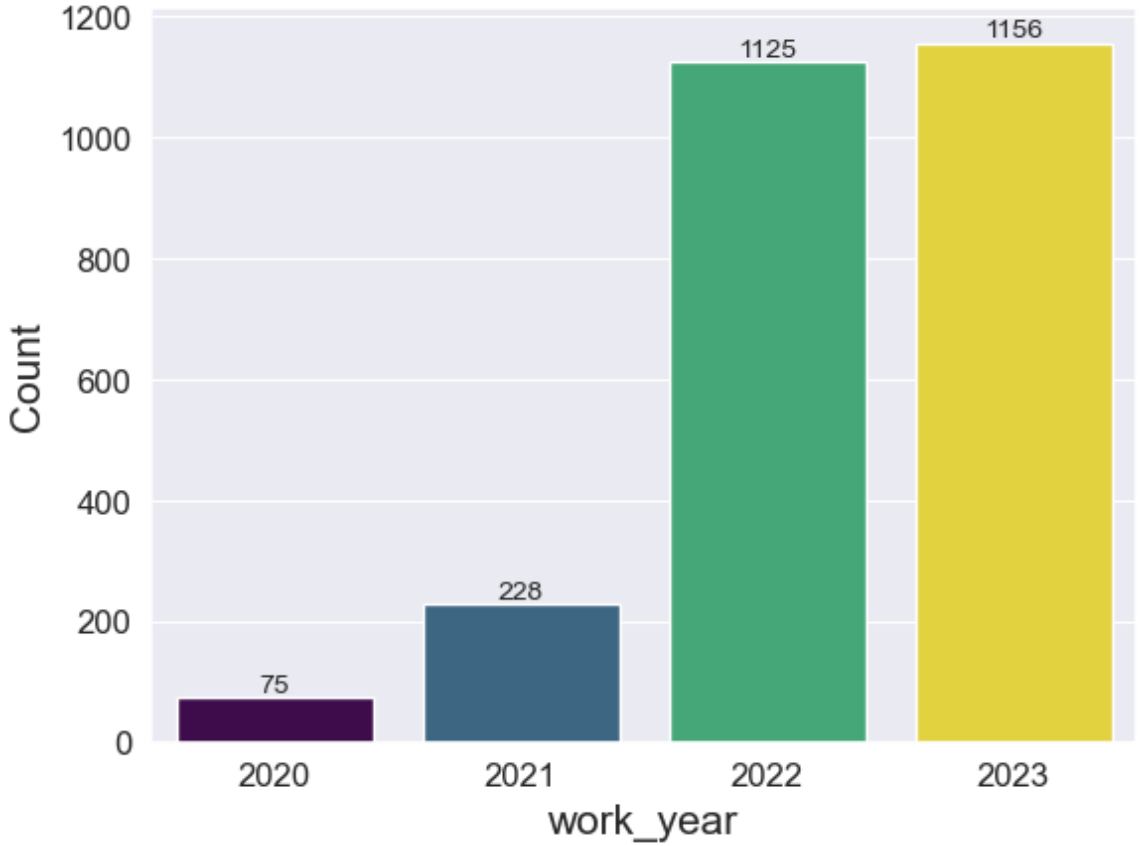
Salary Segmentation

- Entry level **< 120k**
- Intermediate **120k ~ 200k**
- High level **> 200k**

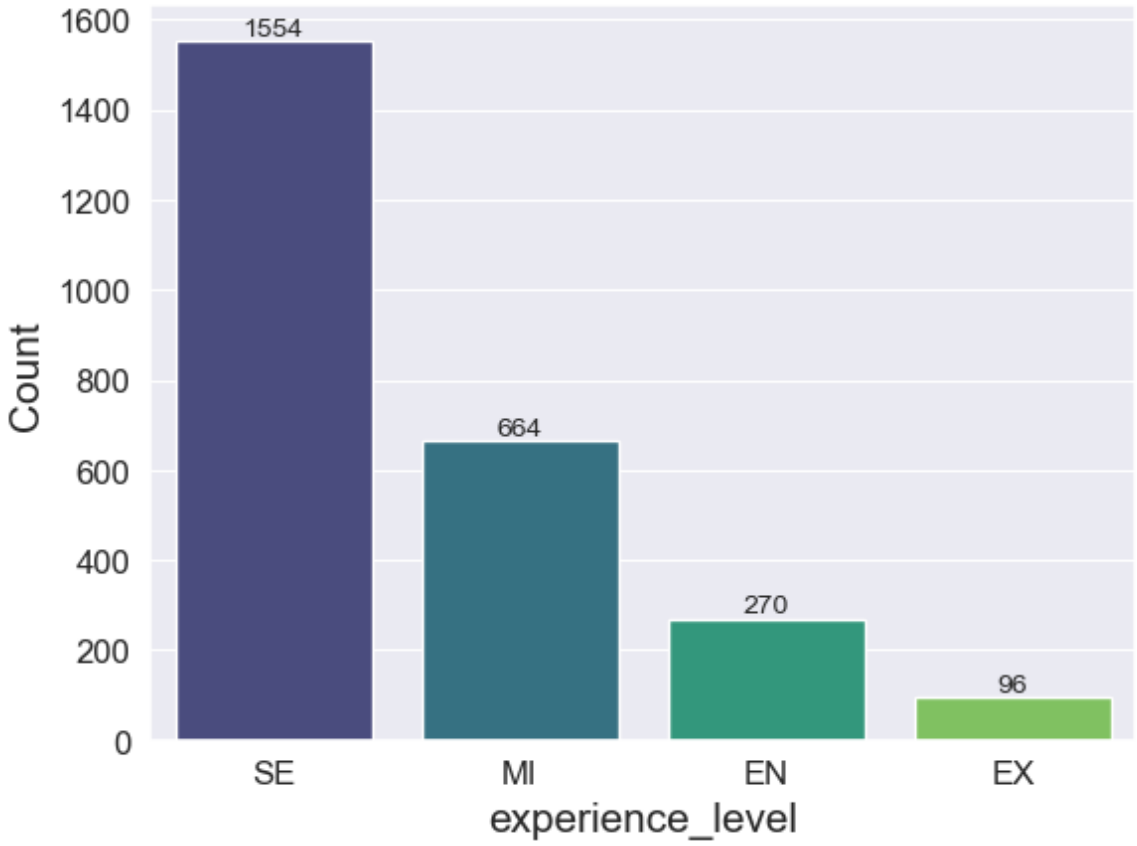


CATEGORICAL VARIABLES

WORK YEAR



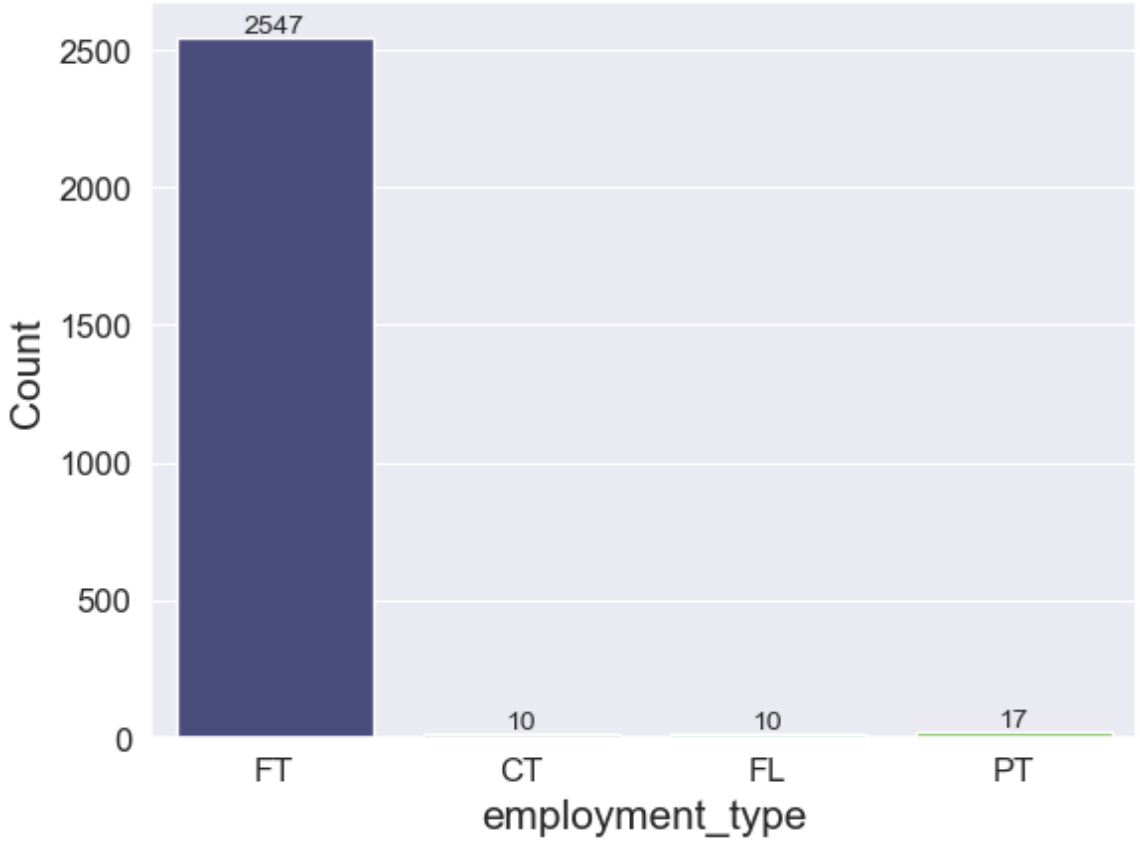
EXPERIENCE LEVEL



CATEGORICAL VARIABLES

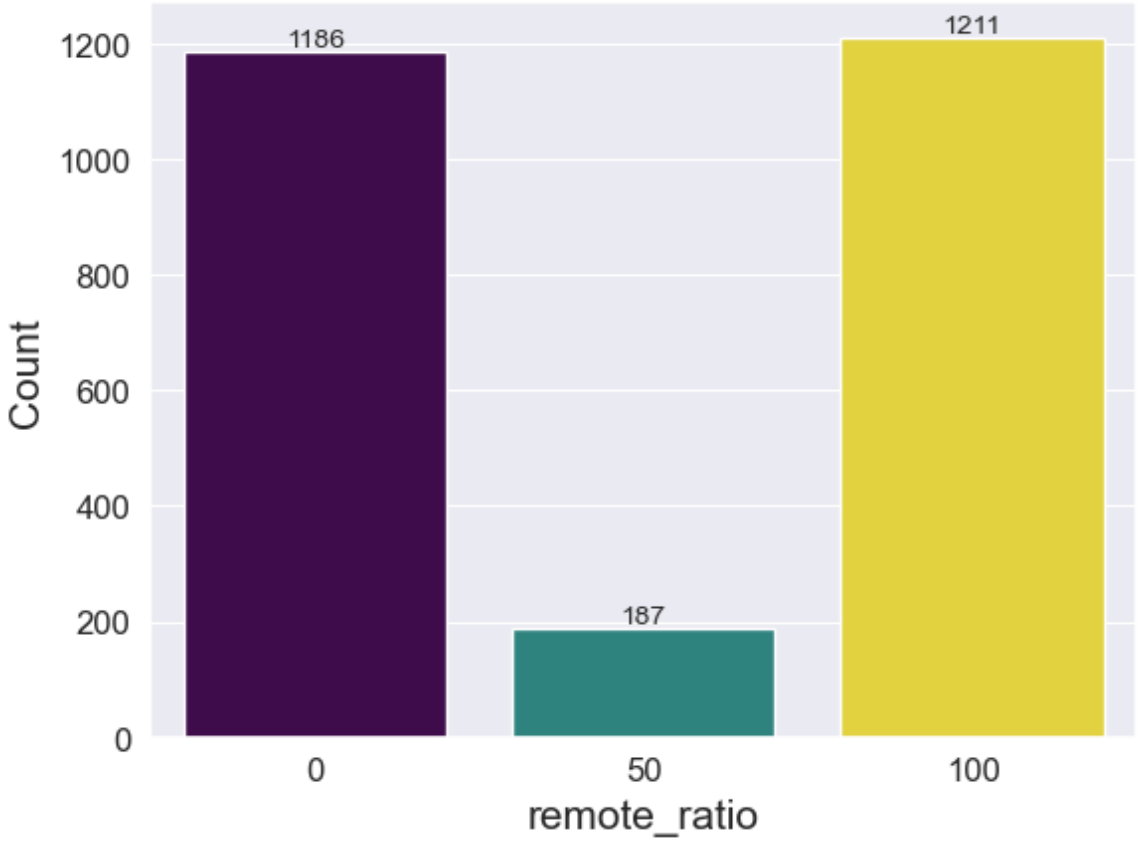
EMPLOYMENT TYPE

1: full time 0: not full time



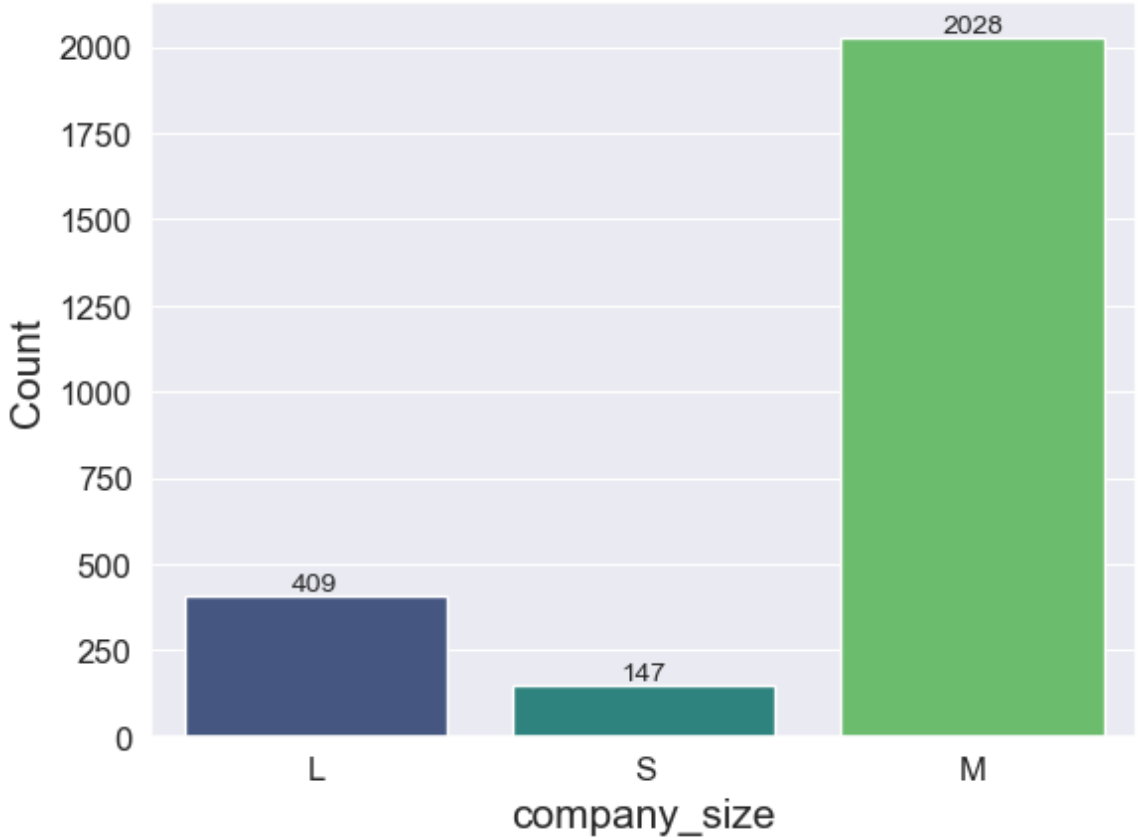
REMOTE RATIO

0: office 50: hybrid 100: home

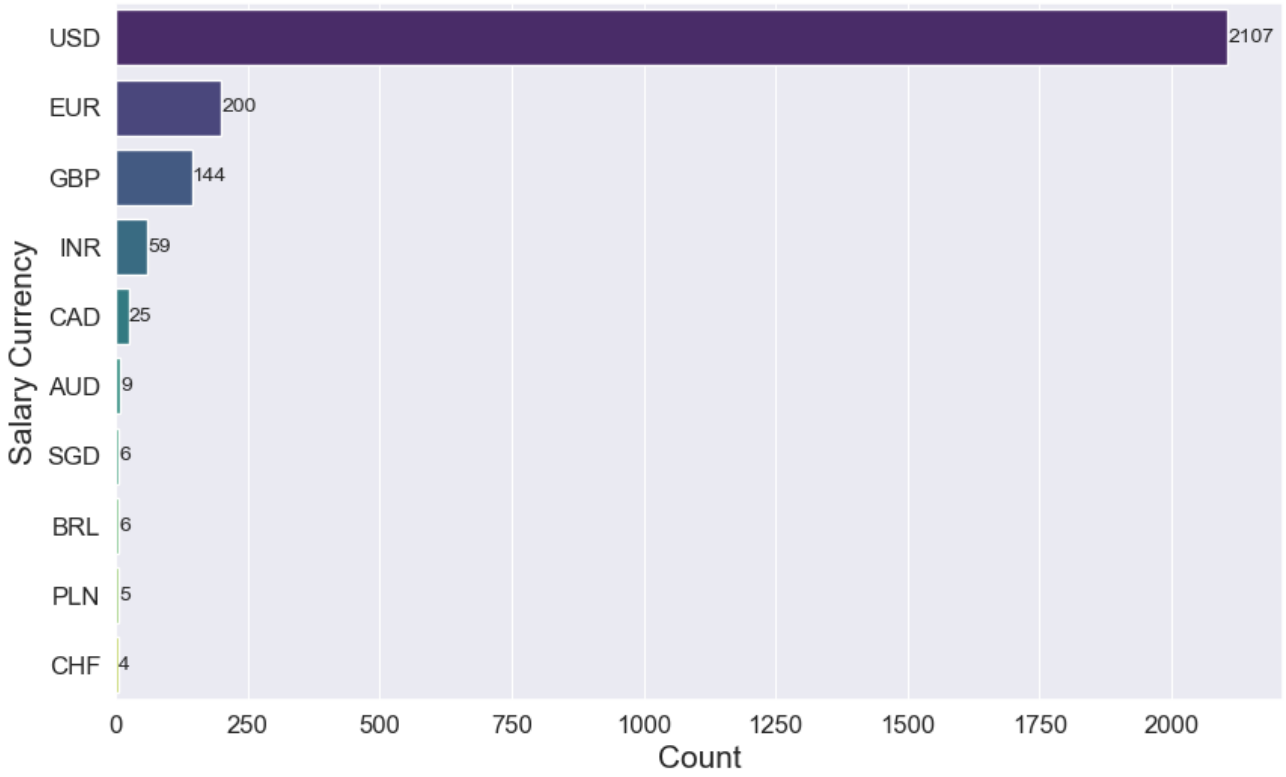


CATEGORICAL VARIABLES

COMPANY SIZE



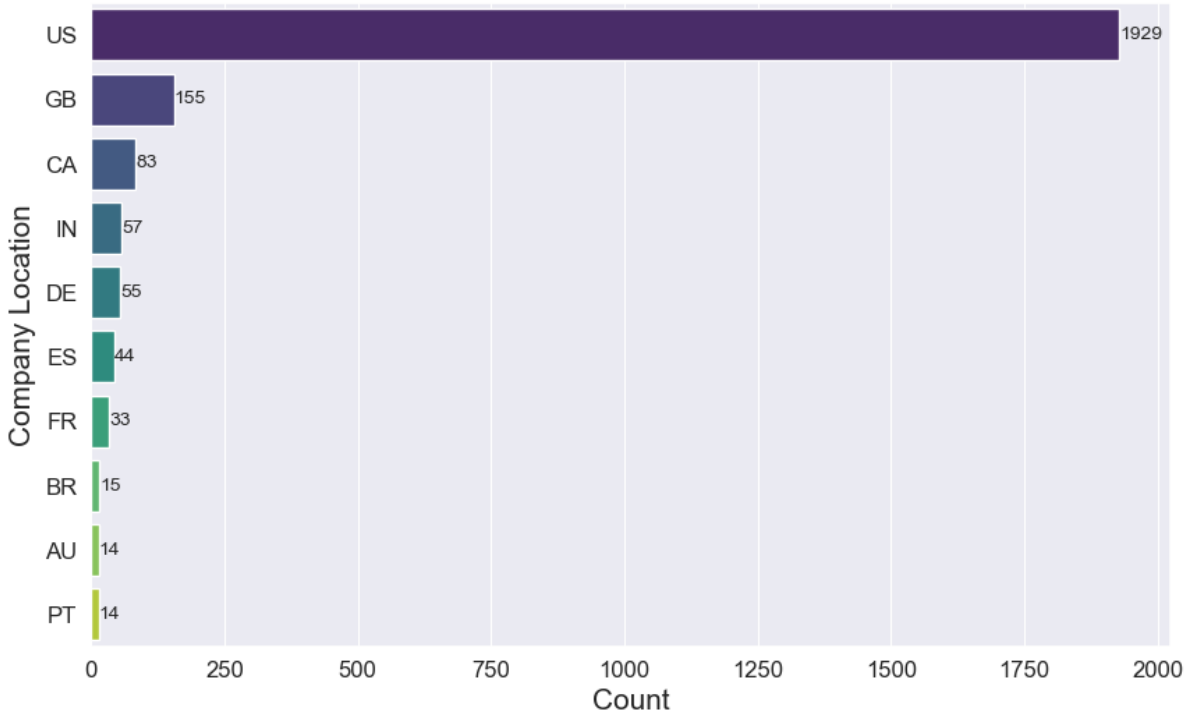
SALARY CURRENCY



CATEGORICAL VARIABLES

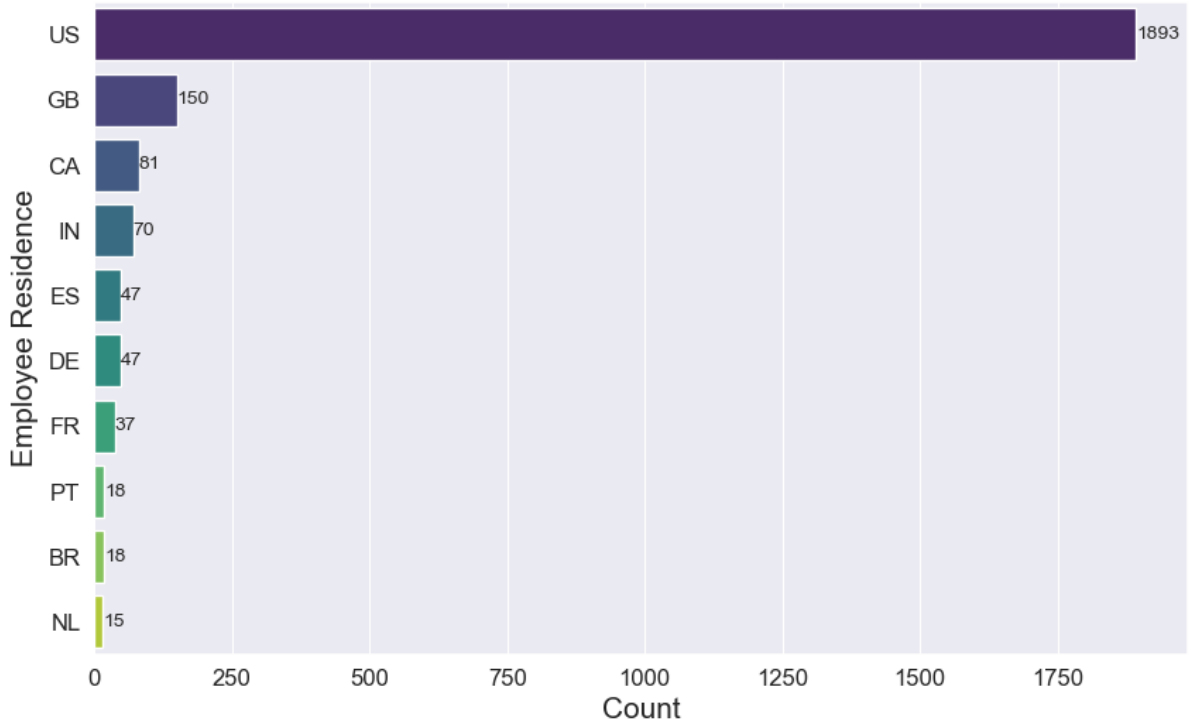
COMPANY LOCATION

0: not in US 1: in US

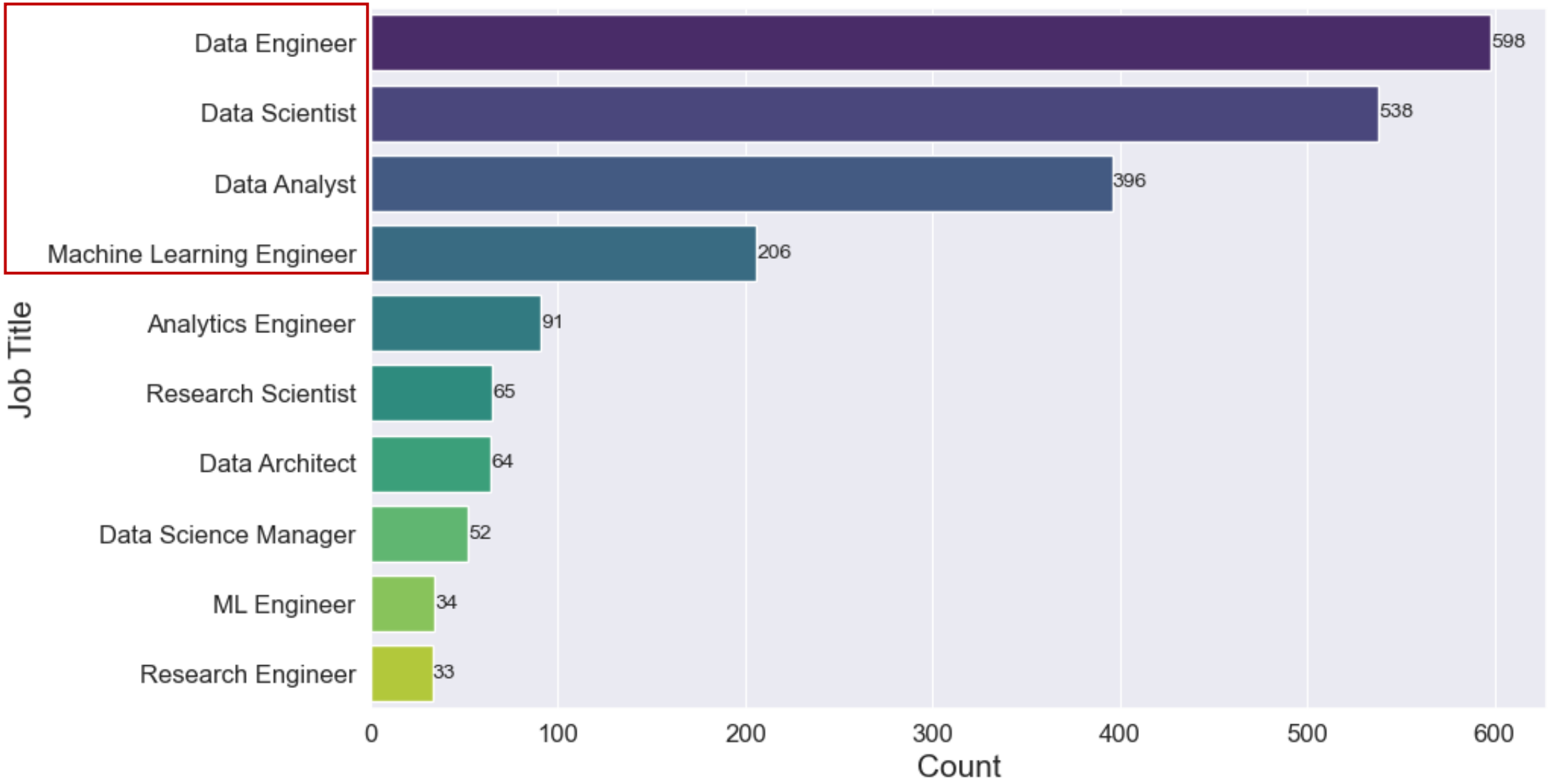


EMPLOYEE RESIDENCE

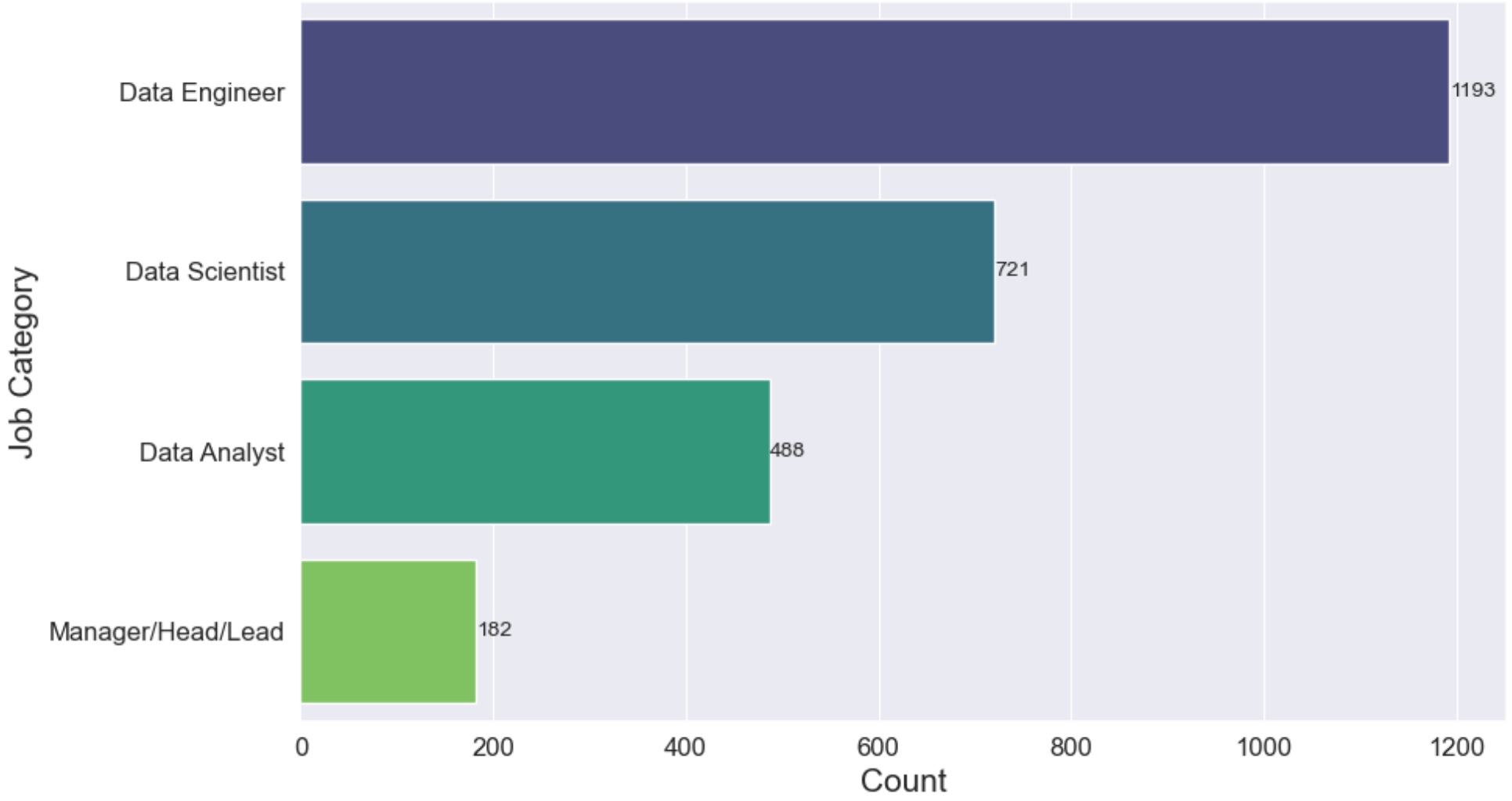
0: not in the US 1: US



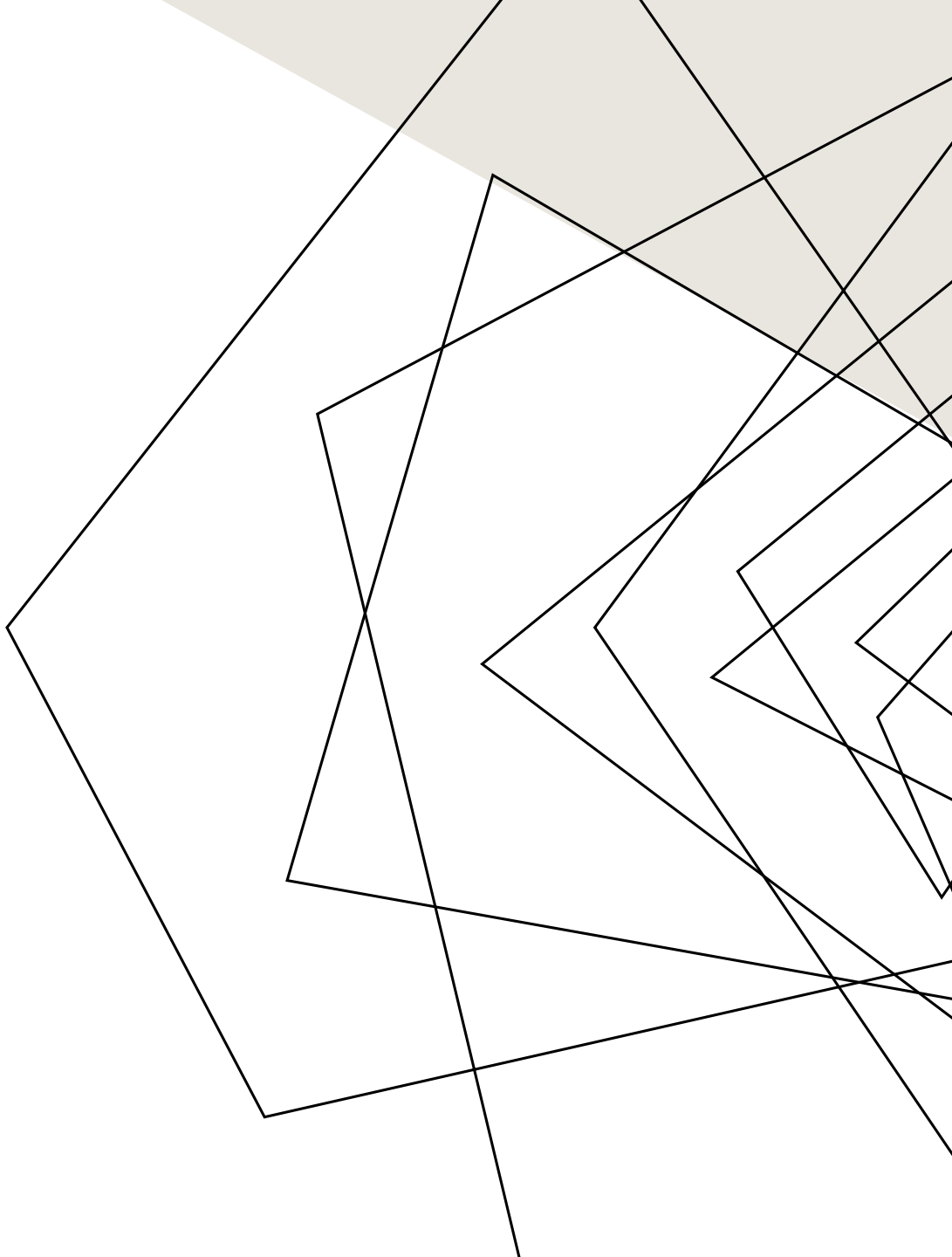
TOP 10 JOB TITLES



JOB CATEGORY



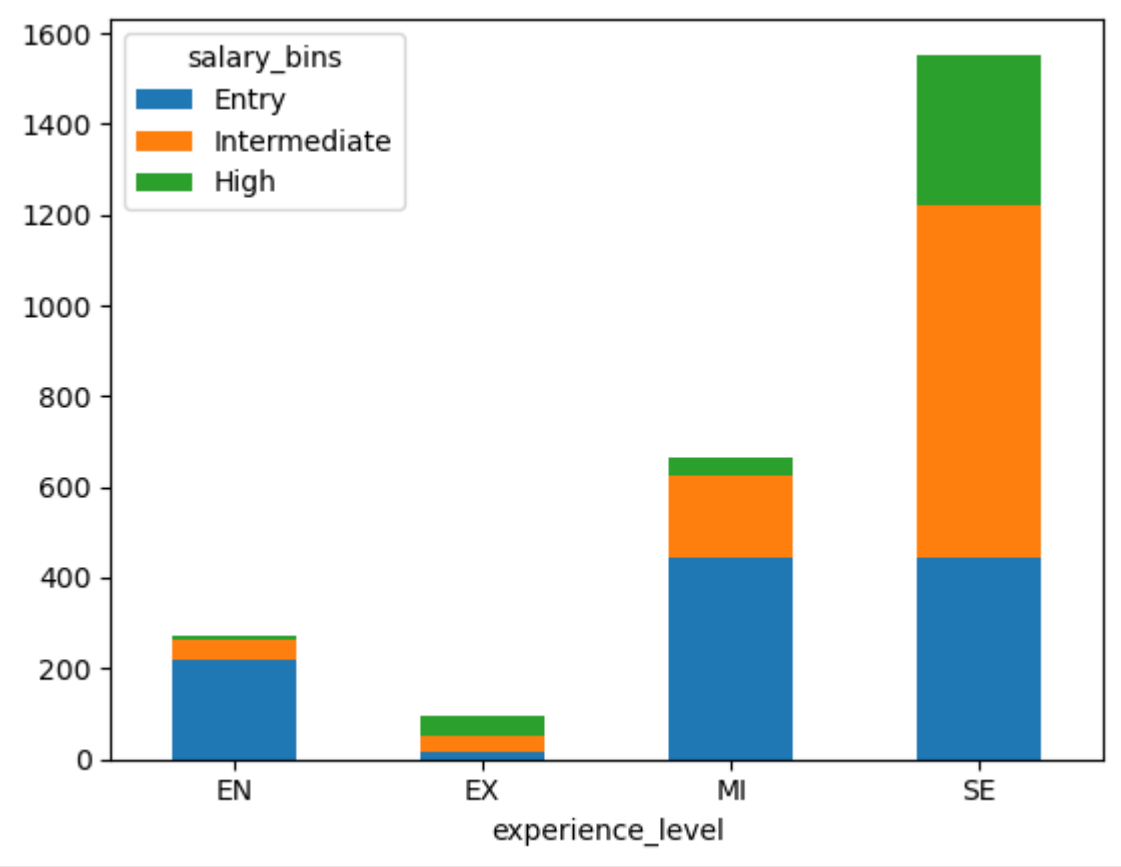
BIVARIATE ANALYSIS

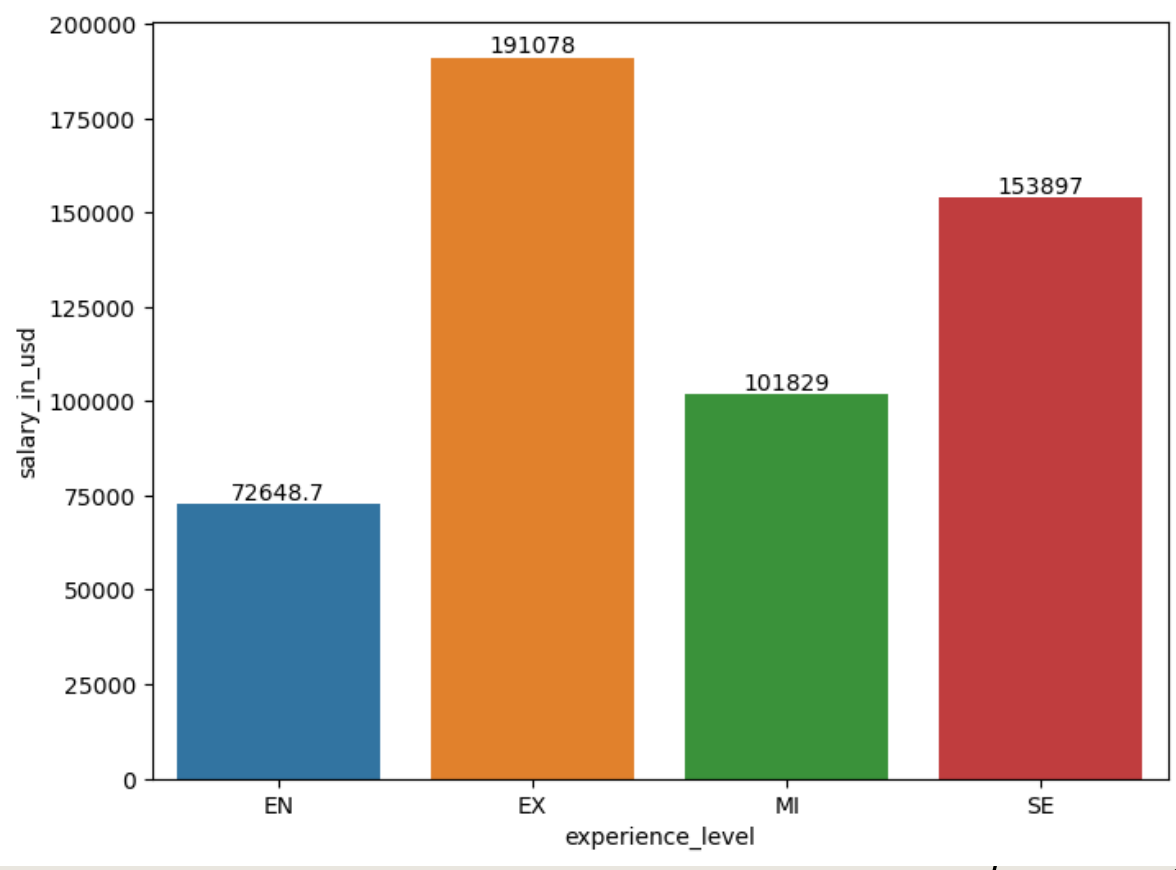


CHI-SQUARE TEST

work_year | **experience_level** | full_time | employee_residence(US) | remote | company_location(US) | company_size | job_category

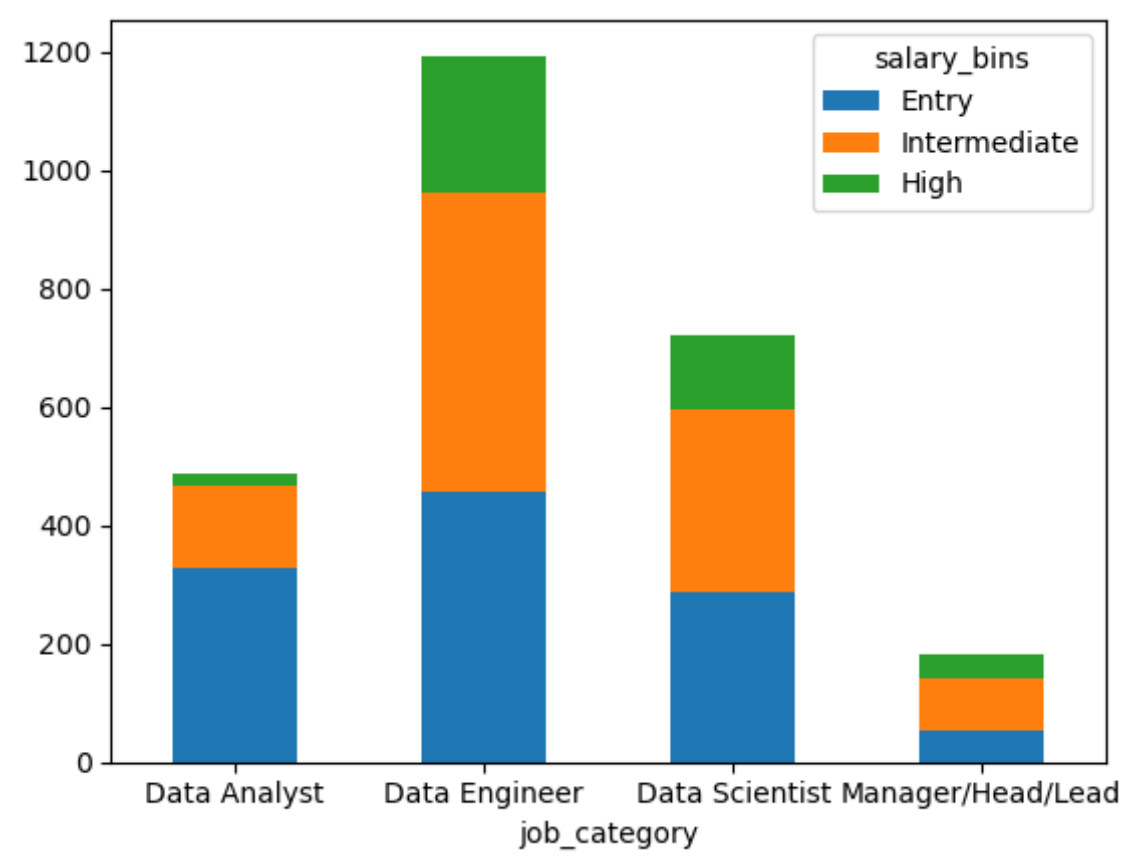
All the categorical variables are significantly associated with **salary_bins** at the 5% significance level.

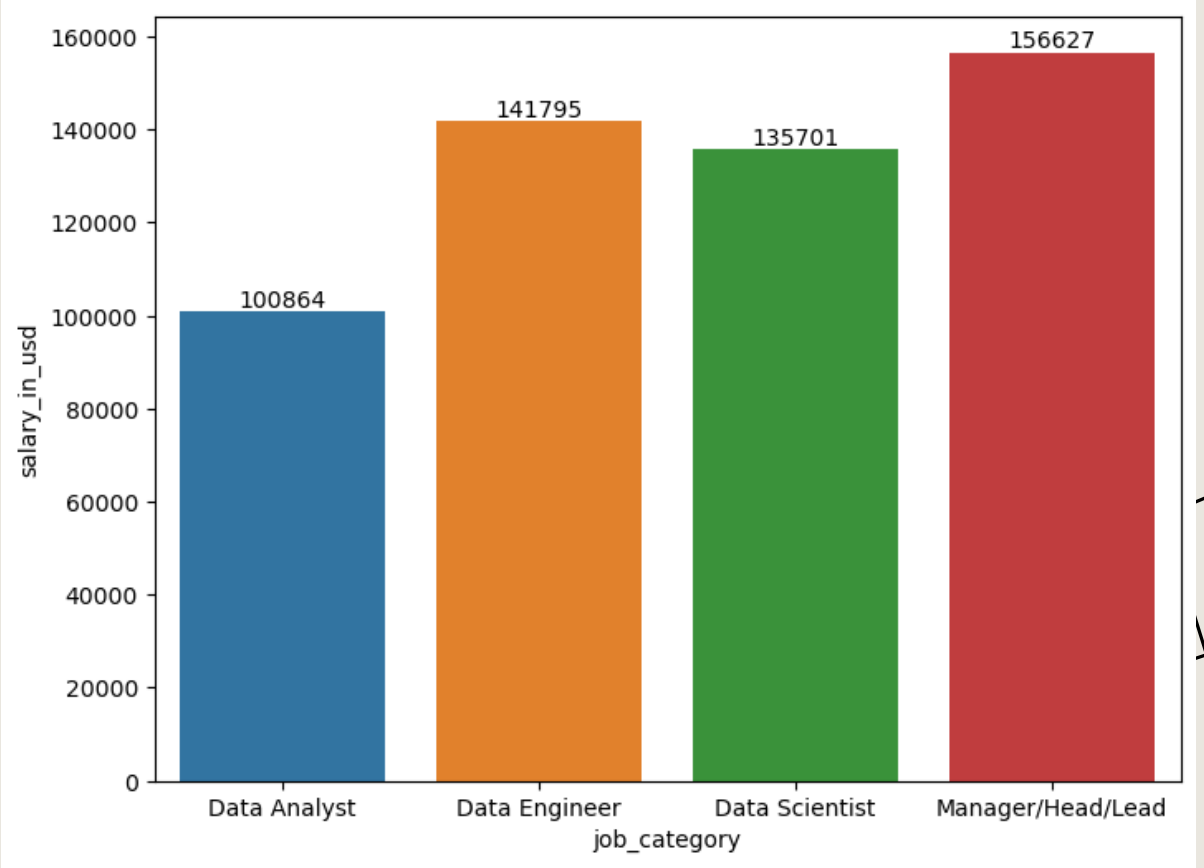




CHI-SQUARE TEST

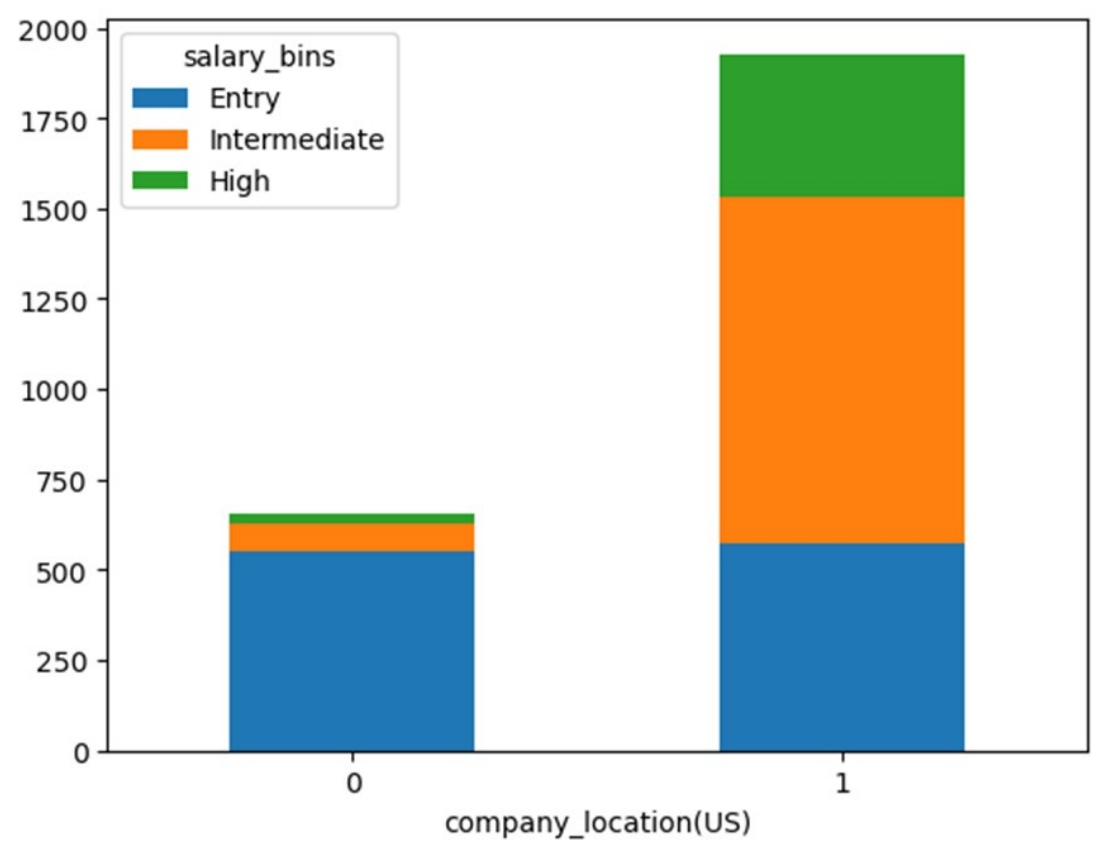
work_year | experience_level | full_time | employee_residence(US) | remote | company_location(US) | company_size | **job_category**

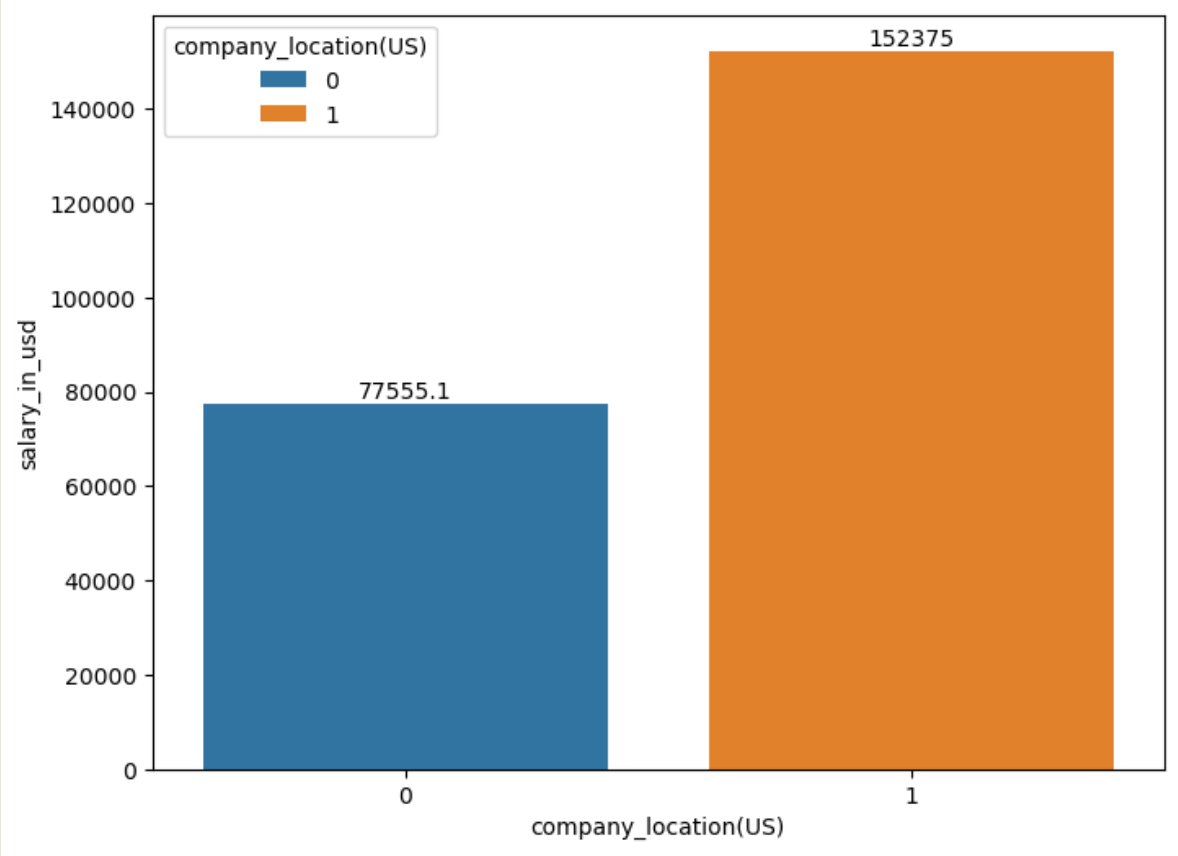




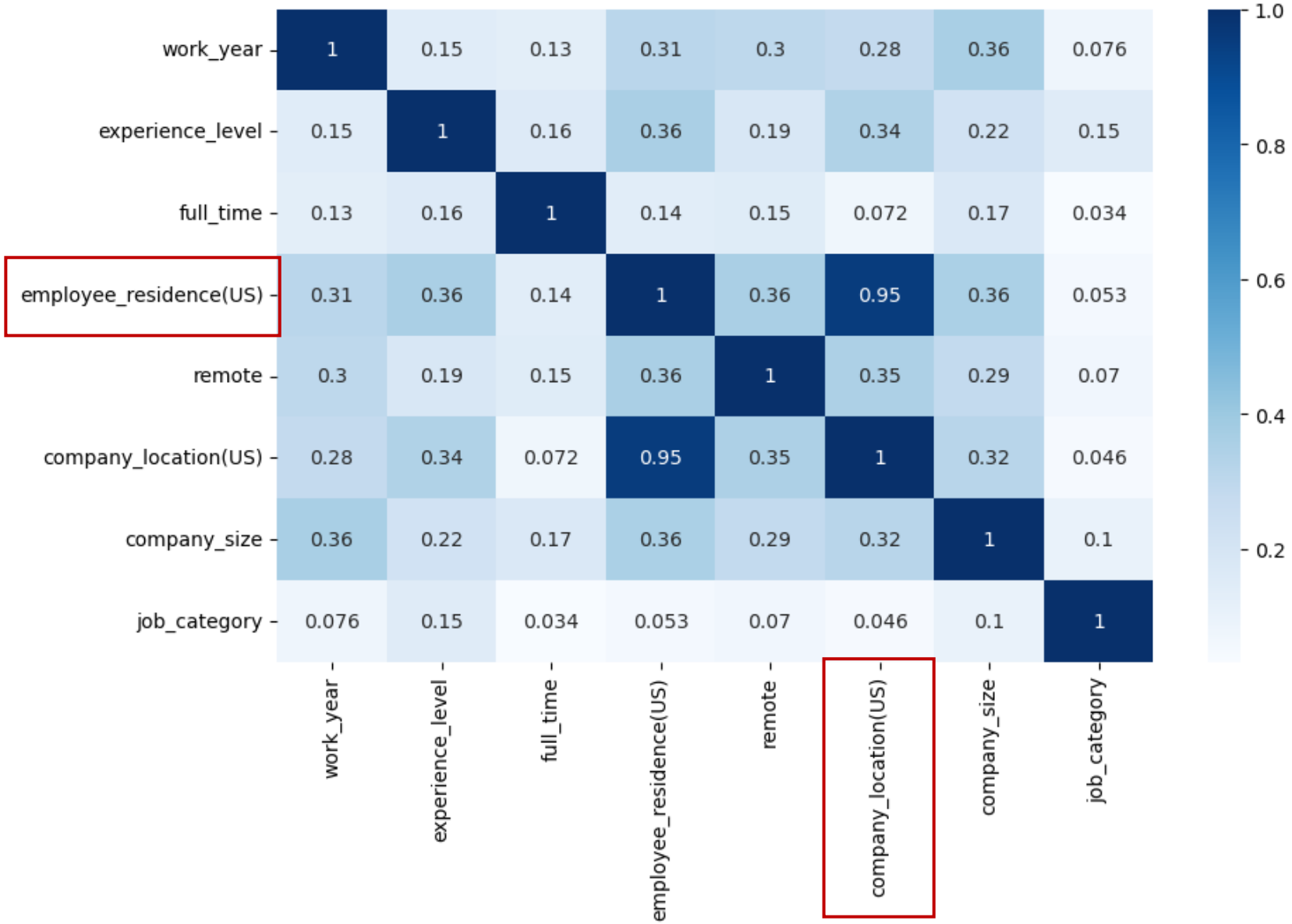
CHI-SQUARE TEST

work_year | experience_level | full_time | employee_residence(US) | remote | **company_location(US)** | company_size | job_category



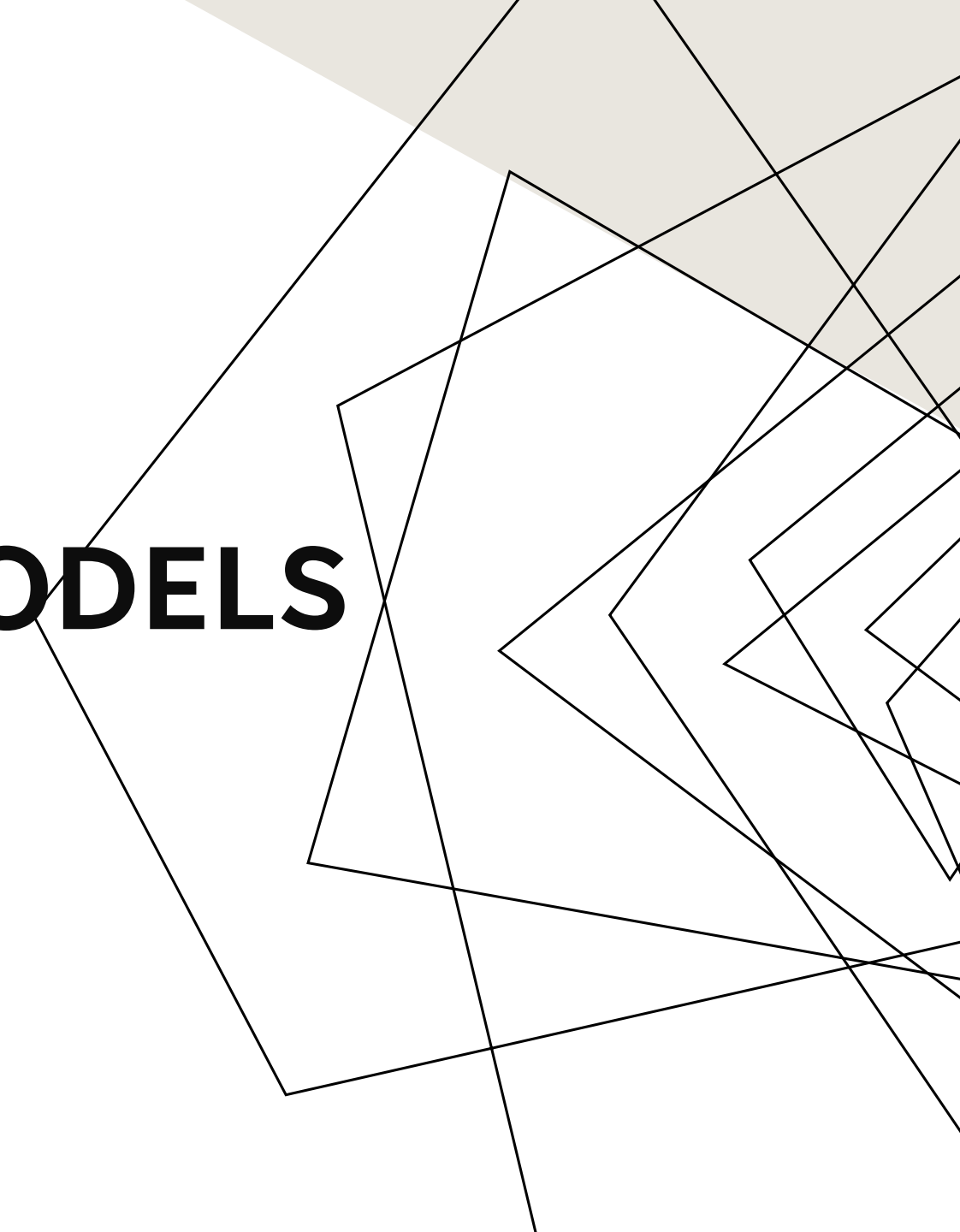


CRAMER'S V HEATMAP



CLASSIFICATION MODELS

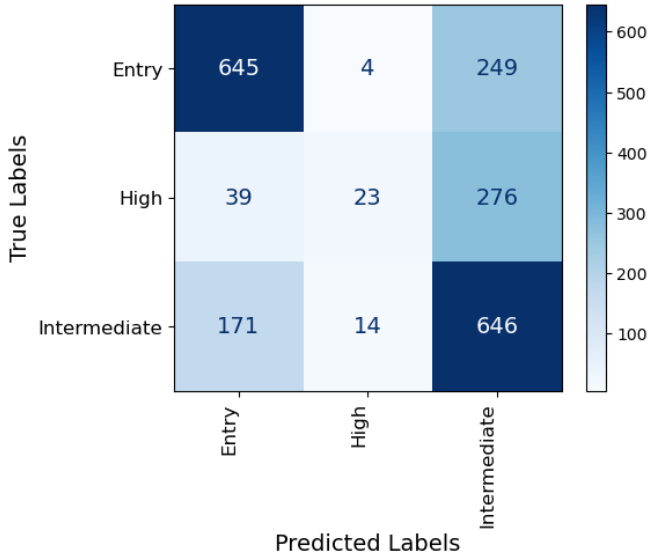
LOGISTIC REGRESSION



DEFAULT PARAMETERS

Training

- Cross-Val : 0.6202

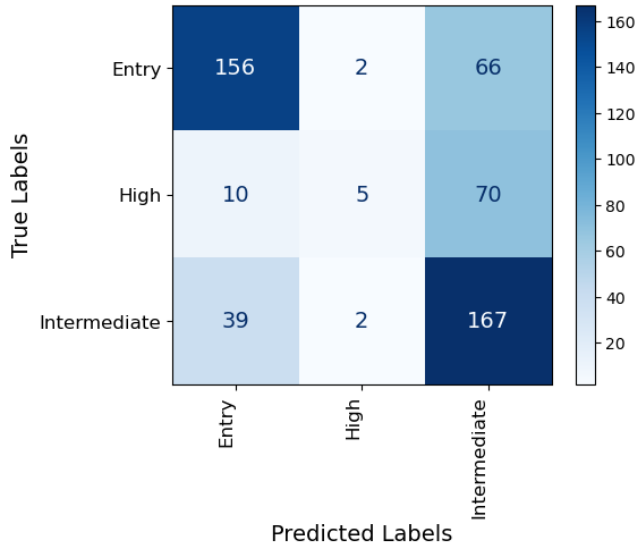


Classification Report:

	precision	recall	f1-score	support
Entry	0.75	0.72	0.74	898
High	0.56	0.07	0.12	338
Intermediate	0.55	0.78	0.65	831
accuracy			0.64	2067
macro avg	0.62	0.52	0.50	2067
weighted avg	0.64	0.64	0.60	2067

Test

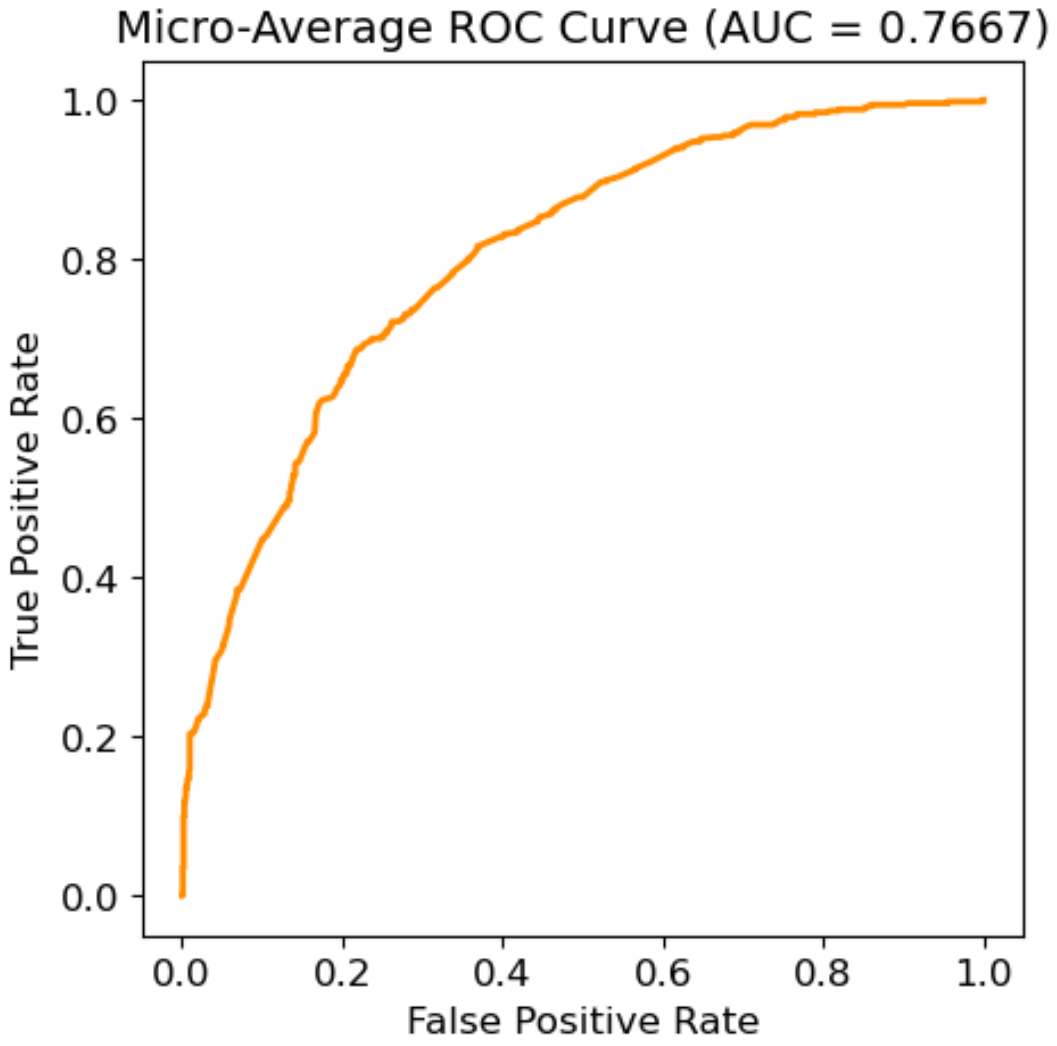
- Accuracy : 0.6344



Classification Report:

	precision	recall	f1-score	support
Entry	0.76	0.70	0.73	224
High	0.56	0.06	0.11	85
Intermediate	0.55	0.80	0.65	208
accuracy			0.63	517
macro avg	0.62	0.52	0.50	517
weighted avg	0.64	0.63	0.60	517

DEFAULT PARAMETERS



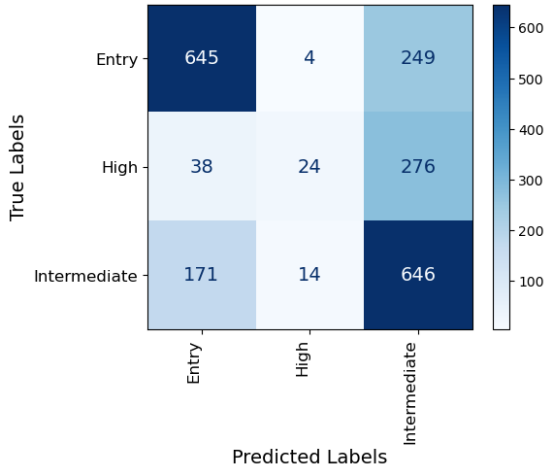
	Entry	Intermediate	High
full_time	-0.217071	0.561686	-0.033341
company_location(US)	-1.003043	0.755436	0.576523
work_year_2021	-0.116093	0.079333	0.093713
work_year_2022	-0.339892	0.251530	0.173271
work_year_2023	-0.452929	0.247460	0.328768
experience_level_EX	-0.502164	0.075879	0.502594
experience_level_MI	-0.210238	0.171040	0.111656
experience_level_SE	-0.809043	0.420285	0.773417
remote_Hybrid	0.069656	-0.014484	-0.121587
remote_Office	-0.017958	-0.028300	0.053534
company_size_M	-0.009890	0.128463	-0.199904
company_size_S	0.189169	-0.122323	-0.153838
job_category_Data Engineer	-0.617747	0.244889	0.679747
job_category_Data Scientist	-0.666898	0.317689	0.596462
job_category_Manager/Head/Lead	-0.444306	0.242178	0.353945

GRID SEARCH: BEST PARAMETERS

```
Time taken by Grid Search over 252 combinations of hyperparameters 18.472650051116943
Best Parameters: {'C': 2, 'class_weight': None, 'penalty': 'l2', 'solver': 'lbfgs'}
Best Cross-Validation Score: 0.6216736264636042
```

Training

- Cross-Val : 0.6216



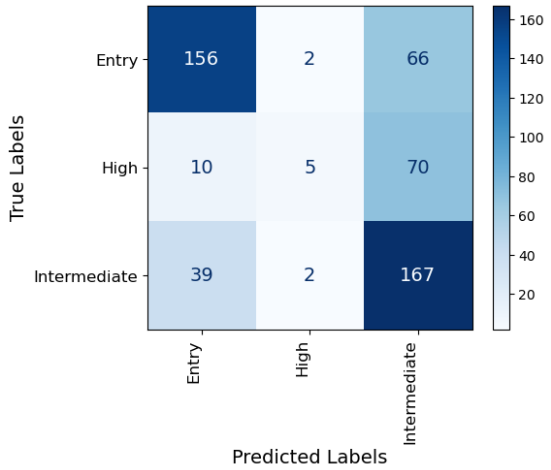

```
Classification Report:
                precision    recall  f1-score   support

         Entry       0.76      0.72      0.74       898
          High       0.57      0.07      0.13       338
  Intermediate       0.55      0.78      0.65       831

      accuracy                           0.64      2067
     macro avg       0.63      0.52      0.50      2067
  weighted avg       0.64      0.64      0.60      2067
```

Test

- Accuracy : 0.6344



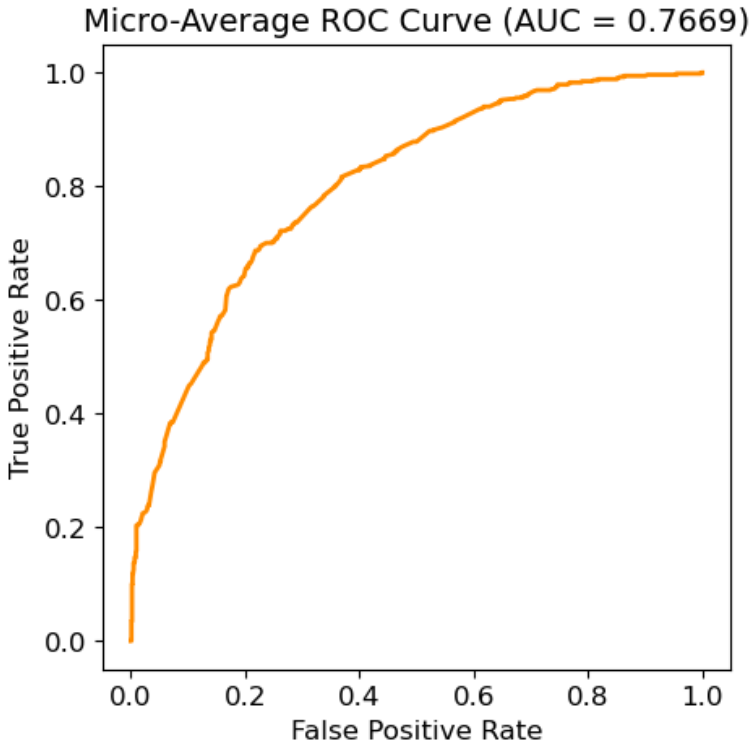

```
Classification Report:
                precision    recall  f1-score   support

         Entry       0.76      0.70      0.73       224
          High       0.56      0.06      0.11        85
  Intermediate       0.55      0.80      0.65       208

      accuracy                           0.63       517
     macro avg       0.62      0.52      0.50       517
  weighted avg       0.64      0.63      0.60       517
```

GRID SEARCH: BEST PARAMETERS

Model	Train Cross-Val Score	Test Accuracy Score	Train Roc_Auc	Test Roc_Auc	Train f1 Score	Test f1 Score
LogisticRegression_default	0.620221	0.634429	0.773574	0.766667	0.599001	0.59556
LogisticRegression_tuned	0.621674	0.634429	0.773620	0.766910	0.599992	0.59556

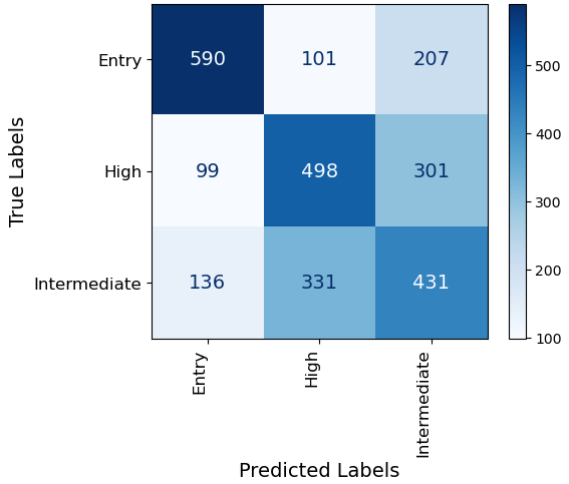


	Entry	Intermediate	High
full_time	-0.218022	0.630163	-0.034830
company_location(US)	-1.005962	0.757201	0.579095
work_year_2021	-0.125651	0.085491	0.106939
work_year_2022	-0.357847	0.262518	0.197784
work_year_2023	-0.471516	0.258703	0.354109
experience_level_EX	-0.505772	0.076888	0.509895
experience_level_MI	-0.216037	0.174081	0.128381
experience_level_SE	-0.815308	0.423141	0.791415
remote_Hybrid	0.069282	-0.013896	-0.121090
remote_Office	-0.017601	-0.028556	0.053186
company_size_M	-0.008119	0.128053	-0.202756
company_size_S	0.189281	-0.122008	-0.154034
job_category_Data Engineer	-0.621480	0.246182	0.689716
job_category_Data Scientist	-0.670867	0.319097	0.605788
job_category_Manager/Head/Lead	-0.446847	0.243210	0.359165

SMOTE - SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

Training

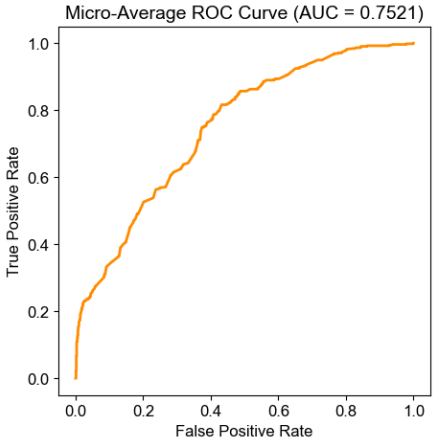
- Cross-Val : 0.5538

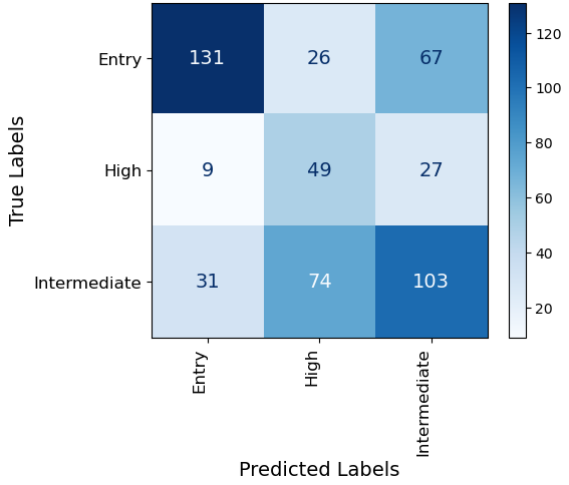


Classification Report:	precision	recall	f1-score	support
Entry	0.72	0.66	0.68	898
High	0.54	0.55	0.54	898
Intermediate	0.46	0.48	0.47	898
accuracy			0.56	2694
macro avg	0.57	0.56	0.57	2694
weighted avg	0.57	0.56	0.57	2694

Test

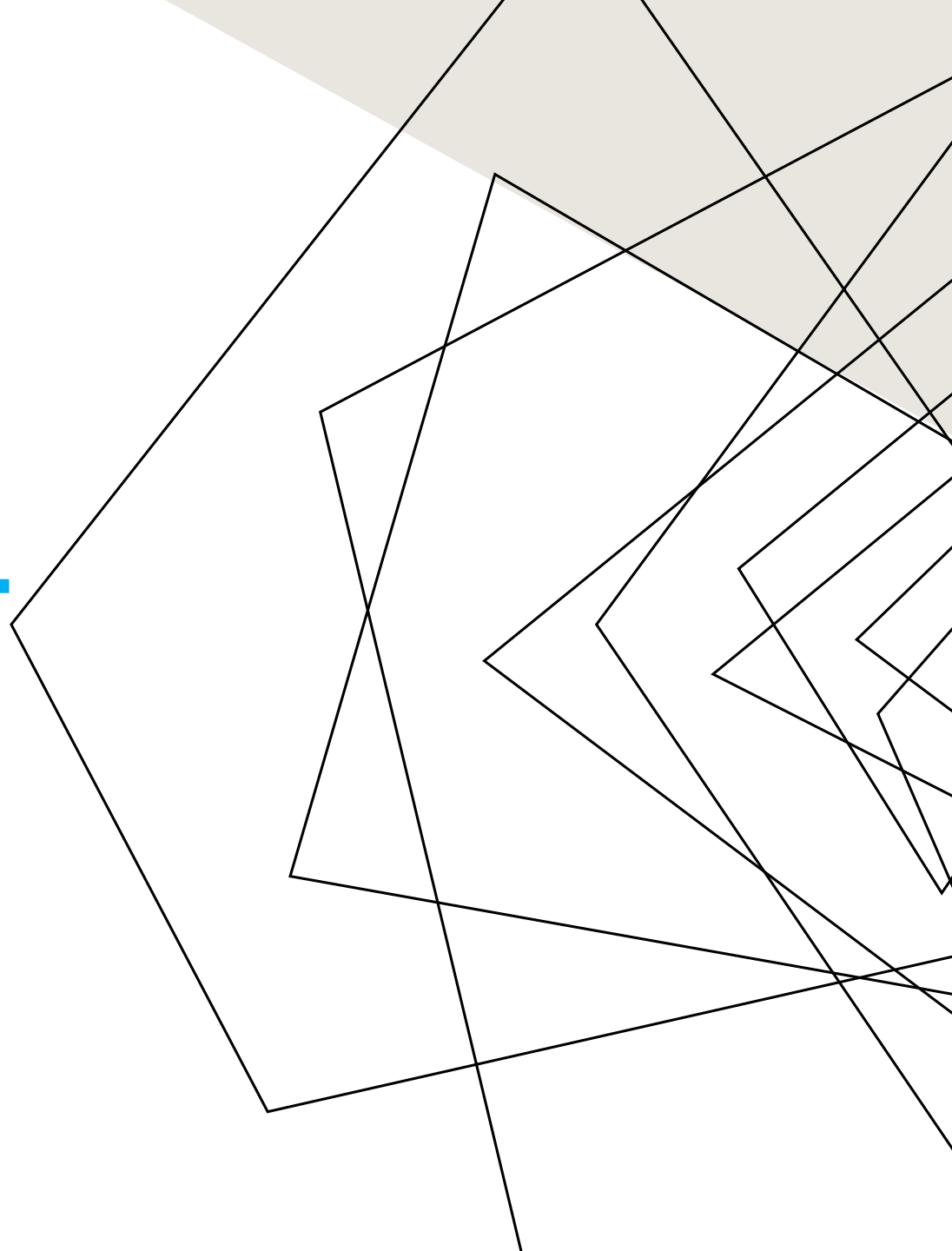
- Accuracy : 0.5473





Classification Report:	precision	recall	f1-score	support
Entry	0.77	0.58	0.66	224
High	0.33	0.58	0.42	85
Intermediate	0.52	0.50	0.51	208
accuracy			0.55	517
macro avg	0.54	0.55	0.53	517
weighted avg	0.60	0.55	0.56	517

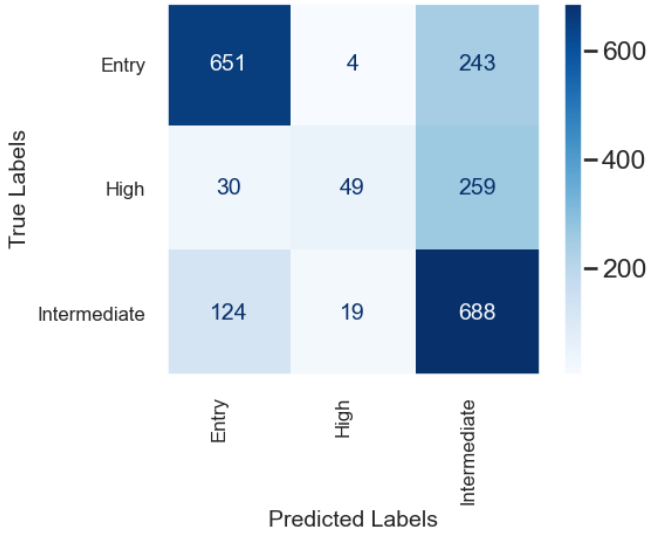
RANDOM FOREST



DEFAULT PARAMETERS

Training

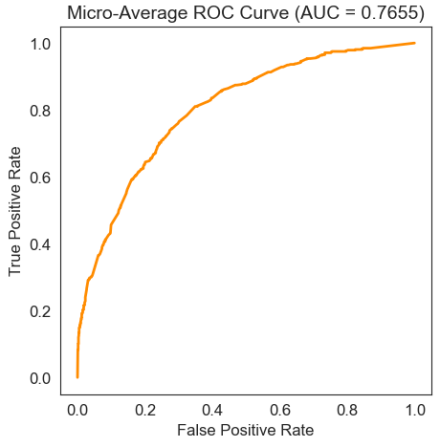
- Cross-Val : 0.6076

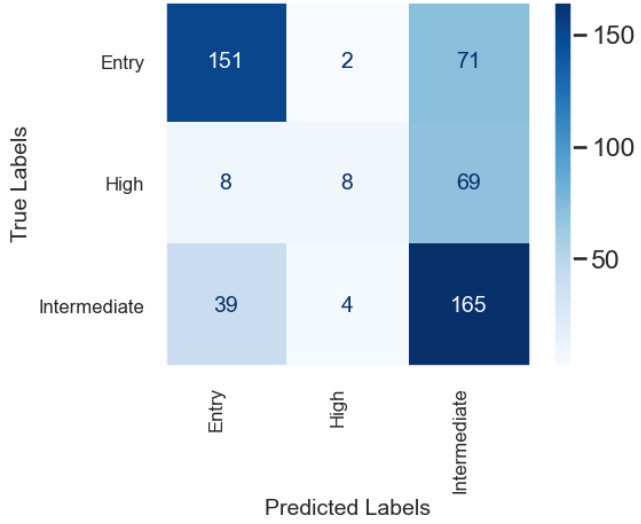


Classification Report:	precision	recall	f1-score	support
Entry	0.81	0.72	0.76	898
High	0.68	0.14	0.24	338
Intermediate	0.58	0.83	0.68	831
accuracy			0.67	2067
macro avg	0.69	0.57	0.56	2067
weighted avg	0.70	0.67	0.64	2067

Test

- Accuracy : 0.6266

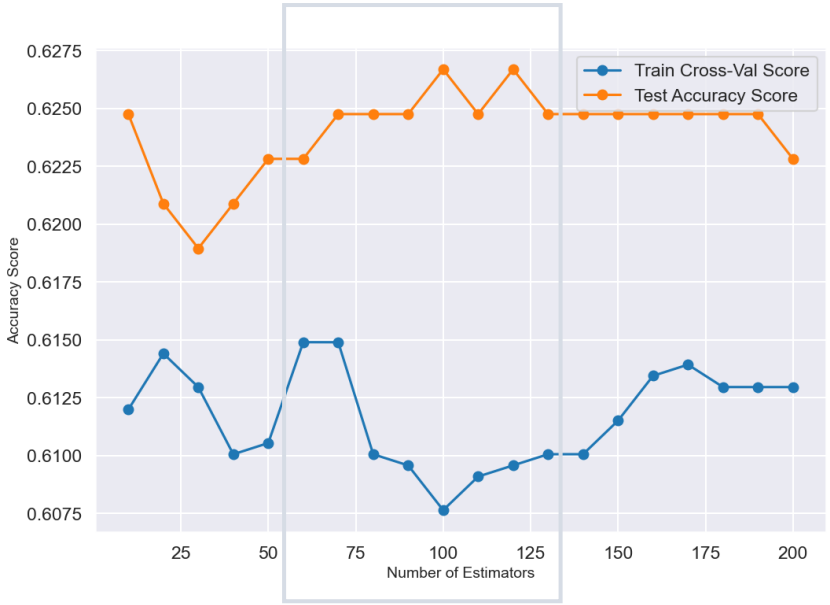


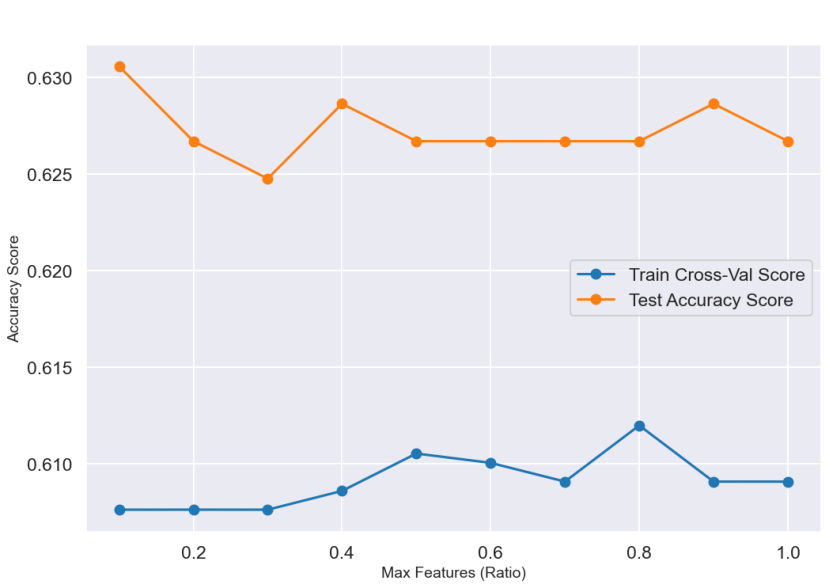


Classification Report:	precision	recall	f1-score	support
Entry	0.76	0.67	0.72	224
High	0.57	0.09	0.16	85
Intermediate	0.54	0.79	0.64	208
accuracy			0.63	517
macro avg	0.63	0.52	0.51	517
weighted avg	0.64	0.63	0.60	517

GRID SEARCH: BEST PARAMETERS

```
Time taken by Grid Search over 5184 combinations of hyperparameters 1304.2928502559662
Best Parameters: {'class_weight': None, 'max_depth': 7, 'max_features': 0.5, 'min_samples_split': 8, 'n_estimators': 110}
Best Cross-Validation Score: 0.6342538980711419
```





Number of Neighbors

60 - 120

Max Depths

7 - 15

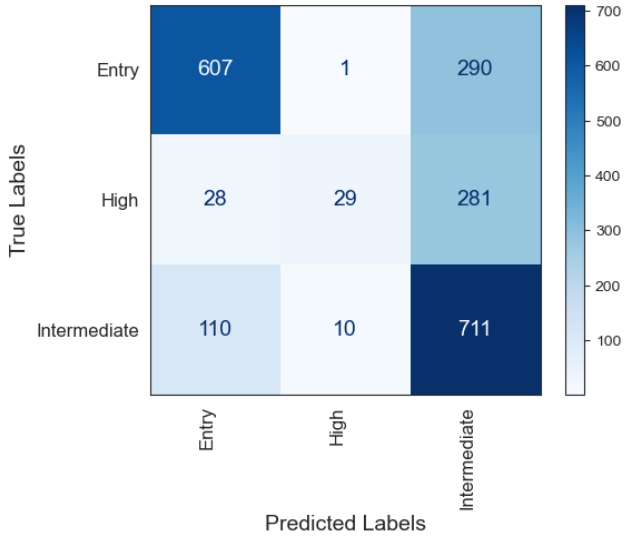
Max Features

0.1 – 0.8

GRID SEARCH: BEST PARAMETERS

Training

- Cross-Val : 0.6342

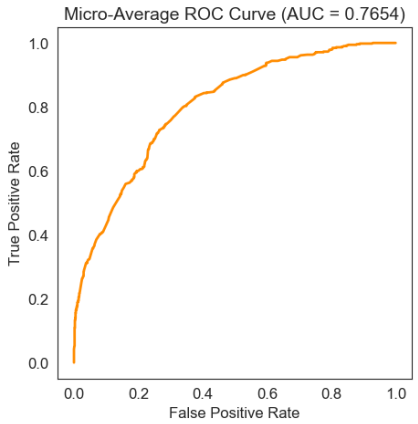


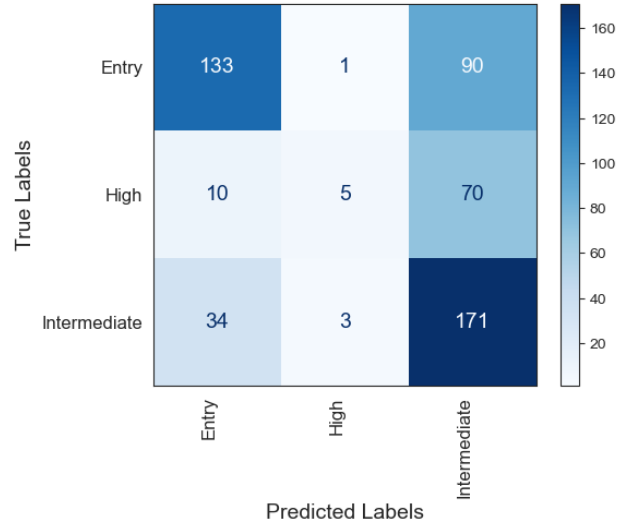
Classification Report:

	precision	recall	f1-score	support
Entry	0.81	0.68	0.74	898
High	0.72	0.09	0.15	338
Intermediate	0.55	0.86	0.67	831
accuracy			0.65	2067
macro avg	0.70	0.54	0.52	2067
weighted avg	0.70	0.65	0.62	2067

Test

- Accuracy : 0.5976

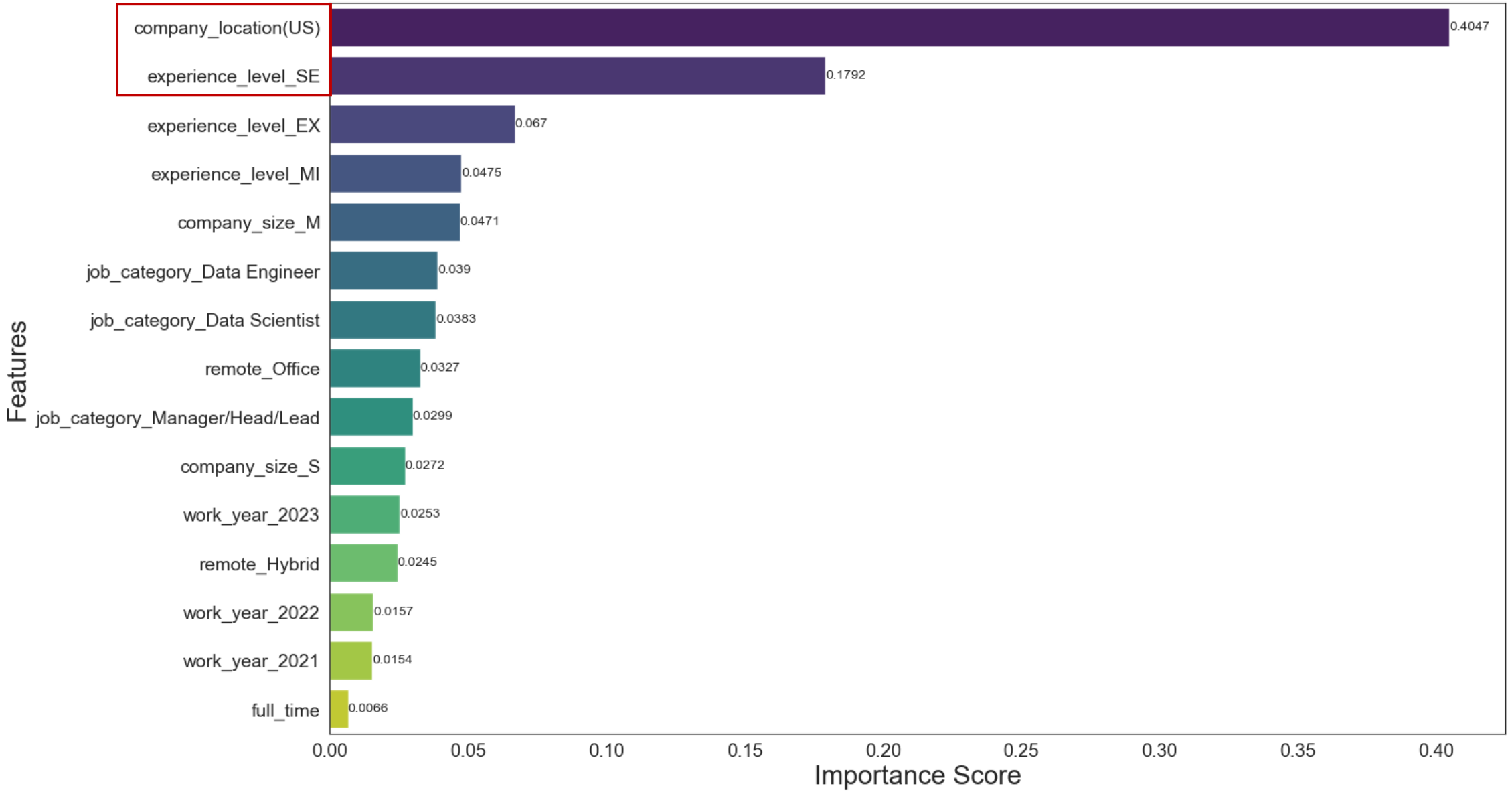




Classification Report:

	precision	recall	f1-score	support
Entry	0.75	0.59	0.66	224
High	0.56	0.06	0.11	85
Intermediate	0.52	0.82	0.63	208
accuracy			0.60	517
macro avg	0.61	0.49	0.47	517
weighted avg	0.62	0.60	0.56	517

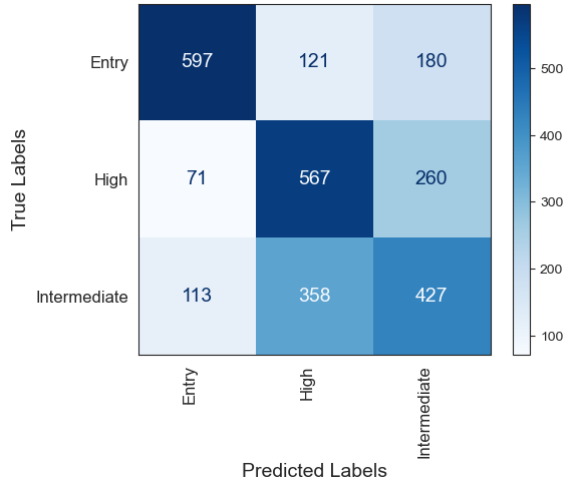
FEATURE IMPORTANCE



SMOTE - SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

Training

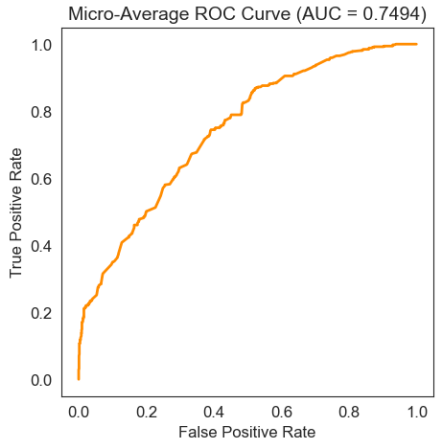
- Cross-Val : 0.5668

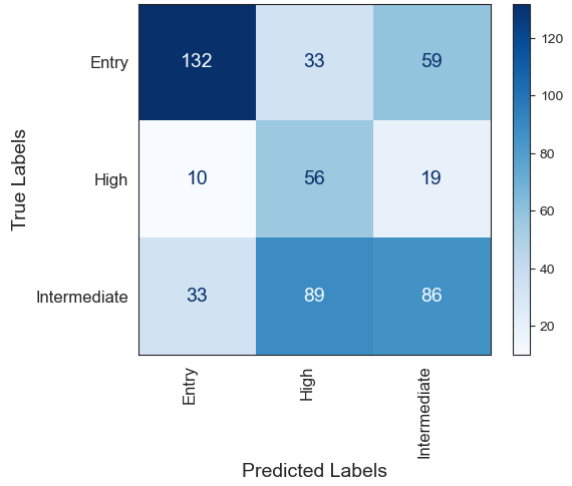


Classification Report:	precision	recall	f1-score	support
Entry	0.76	0.66	0.71	898
High	0.54	0.63	0.58	898
Intermediate	0.49	0.48	0.48	898
accuracy			0.59	2694
macro avg	0.60	0.59	0.59	2694
weighted avg	0.60	0.59	0.59	2694

Test

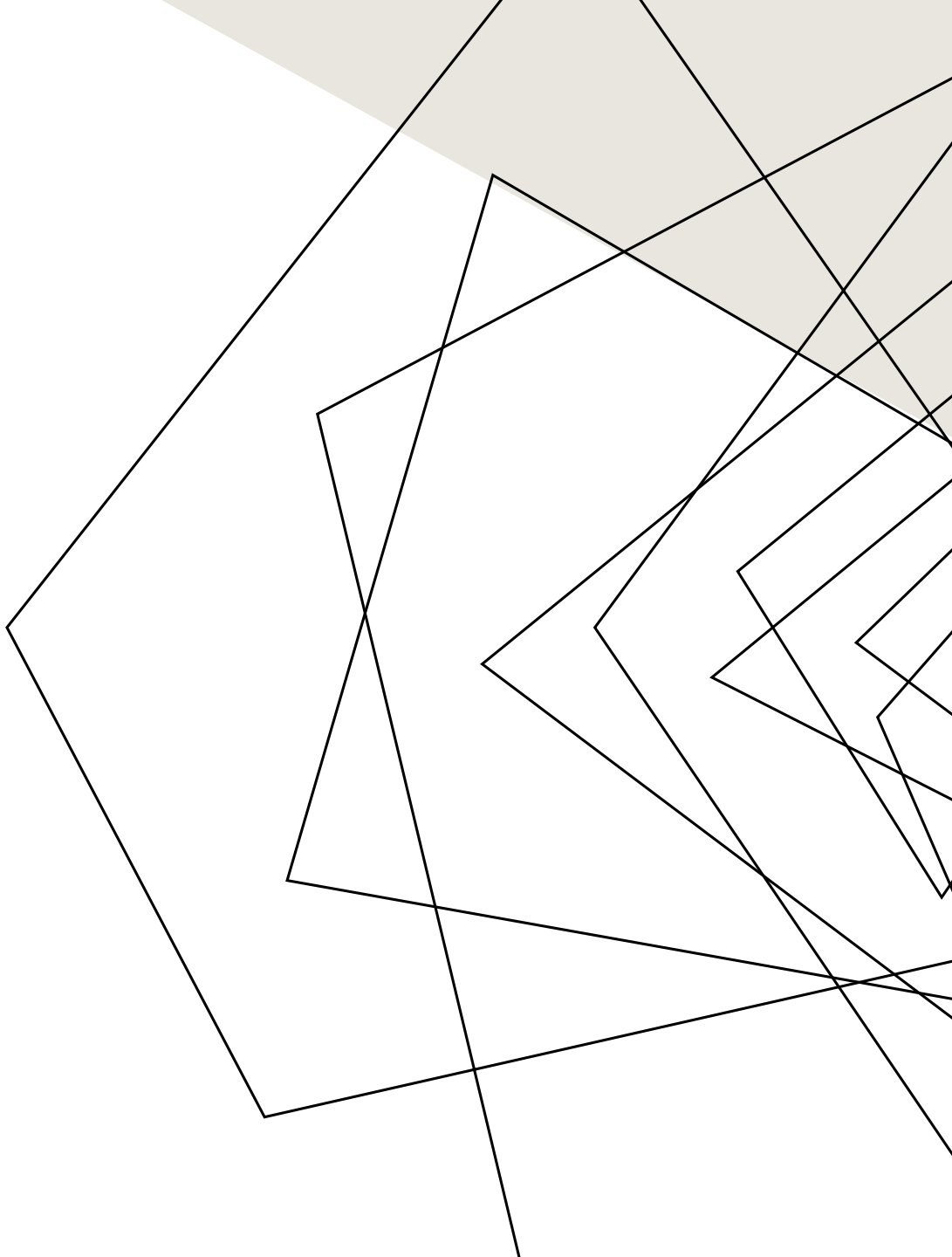
- Accuracy : 0.5299





Classification Report:	precision	recall	f1-score	support
Entry	0.75	0.59	0.66	224
High	0.31	0.66	0.43	85
Intermediate	0.52	0.41	0.46	208
accuracy			0.53	517
macro avg	0.53	0.55	0.52	517
weighted avg	0.59	0.53	0.54	517

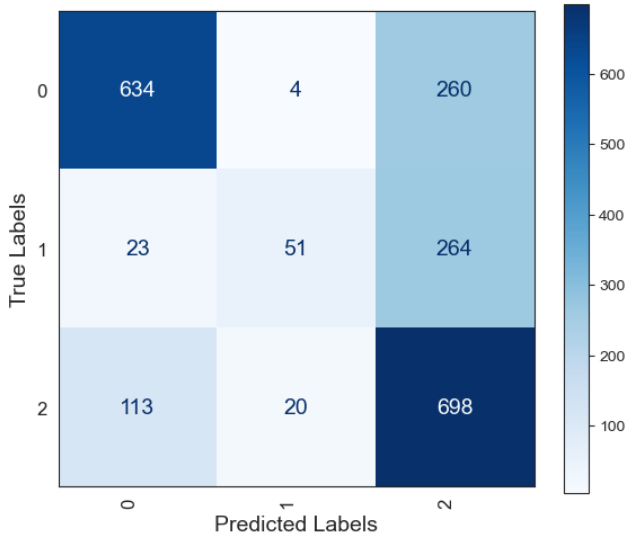
XGBOOST



DEFAULT PARAMETERS

Training

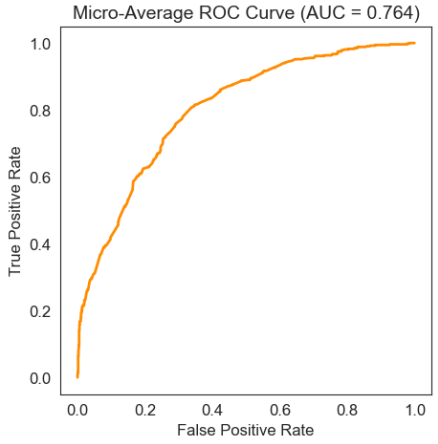
- Cross-Val : 0.6110

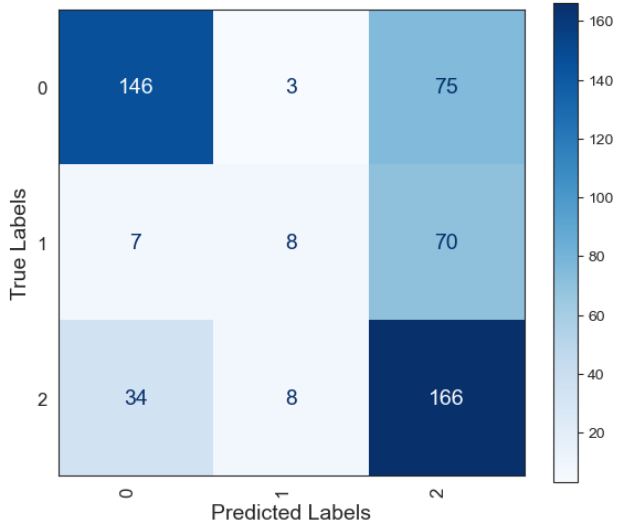


Classification Report:	precision	recall	f1-score	support
0	0.82	0.71	0.76	898
1	0.68	0.15	0.25	338
2	0.57	0.84	0.68	831
accuracy			0.67	2067
macro avg	0.69	0.57	0.56	2067
weighted avg	0.70	0.67	0.64	2067

Test

- Accuracy : 0.6189

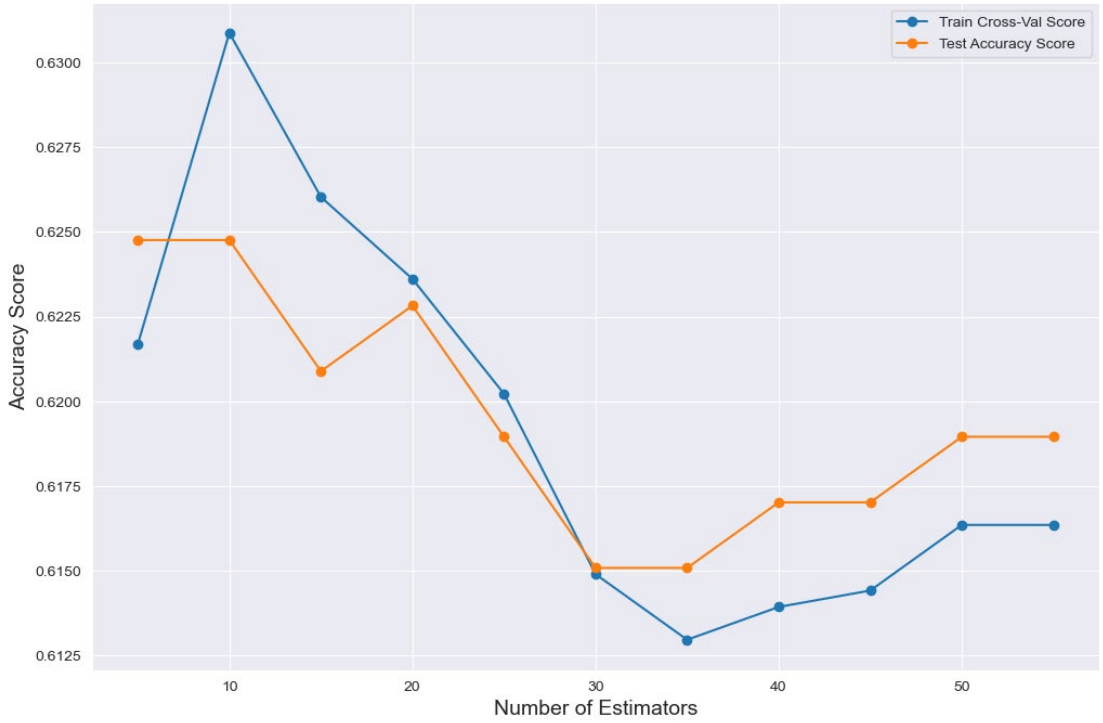




Classification Report:	precision	recall	f1-score	support
0	0.78	0.65	0.71	224
1	0.42	0.09	0.15	85
2	0.53	0.80	0.64	208
accuracy			0.62	517
macro avg	0.58	0.51	0.50	517
weighted avg	0.62	0.62	0.59	517

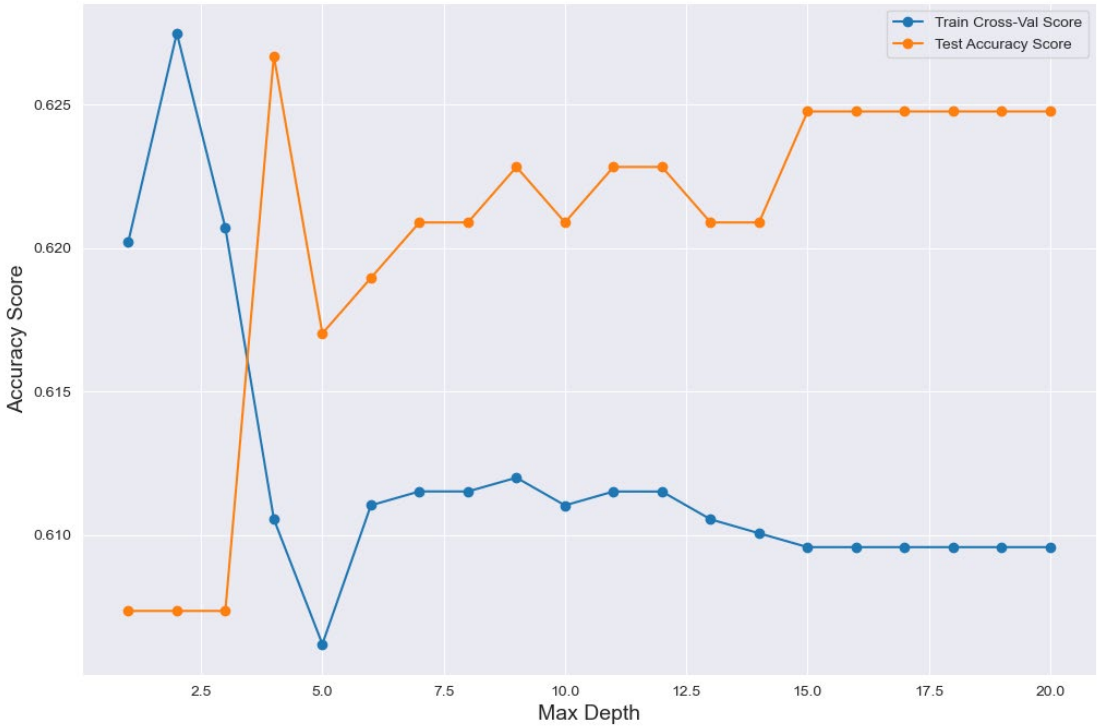
GRID SEARCH: BEST PARAMETERS

```
Time taken by Grid Search over 10560 combinations of hyperparameters 205.7373342514038
Best Parameters: {'booster': 'gbtree', 'colsample_bytree': 0.6, 'gamma': 0, 'learning_rate': 1, 'max_depth': 2, 'n_estimators': 13}
Best Cross-Validation Score: 0.6337684668561604
```



Number of Neighbors

5 - 15



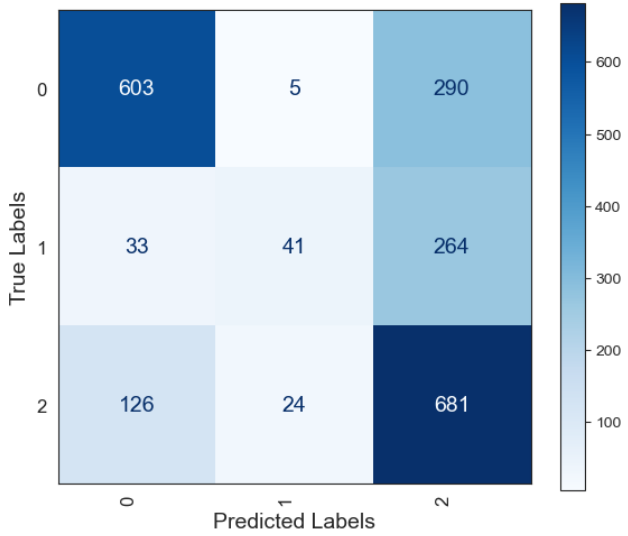
Max Depths

1 - 4

GRID SEARCH: BEST PARAMETERS

Training

- Cross-Val : 0.6337

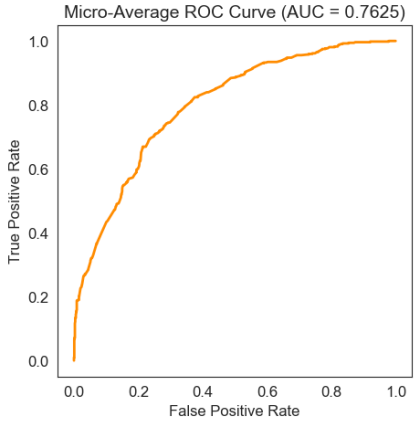


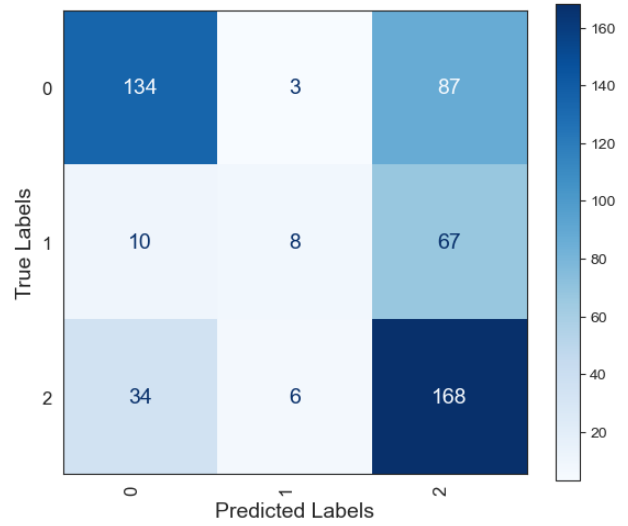
Classification Report:

	precision	recall	f1-score	support
0	0.79	0.67	0.73	898
1	0.59	0.12	0.20	338
2	0.55	0.82	0.66	831
accuracy			0.64	2067
macro avg	0.64	0.54	0.53	2067
weighted avg	0.66	0.64	0.61	2067

Test

- Accuracy : 0.5996

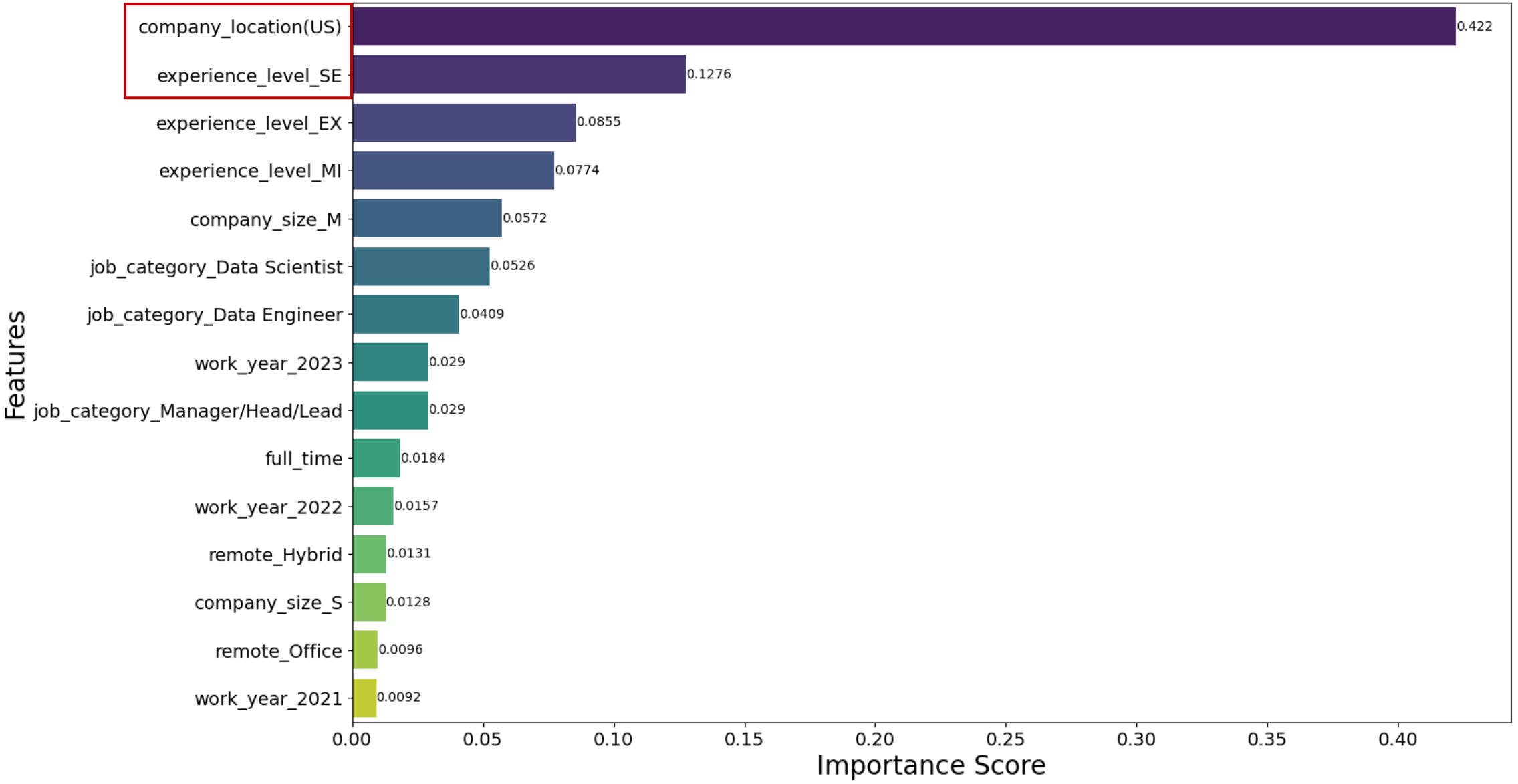




Classification Report:

	precision	recall	f1-score	support
0	0.75	0.60	0.67	224
1	0.47	0.09	0.16	85
2	0.52	0.81	0.63	208
accuracy			0.60	517
macro avg	0.58	0.50	0.49	517
weighted avg	0.61	0.60	0.57	517

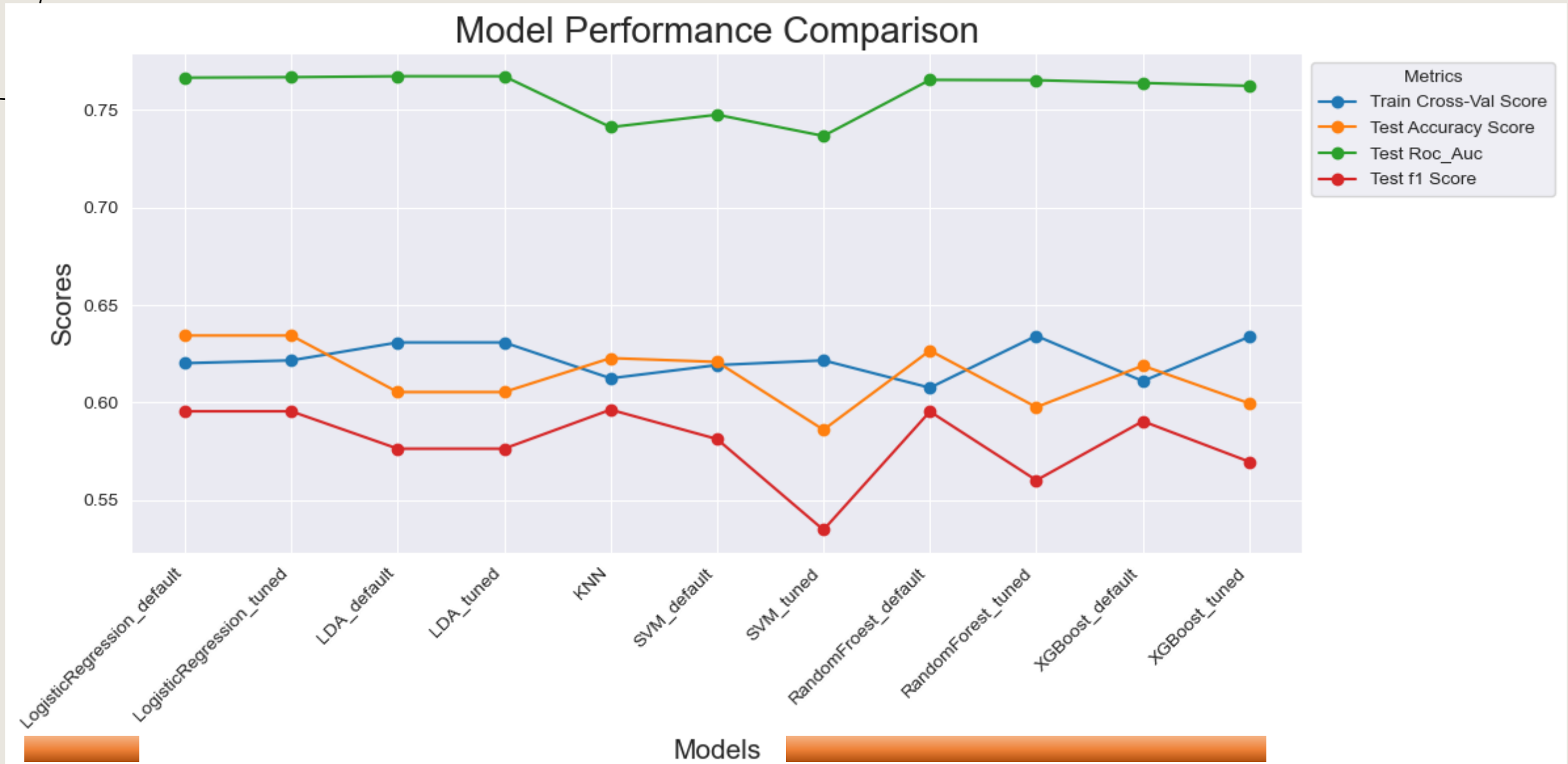
FEATURE IMPORTANCE



PERFORMANCE COMPARISON ACROSS ALL MODELS

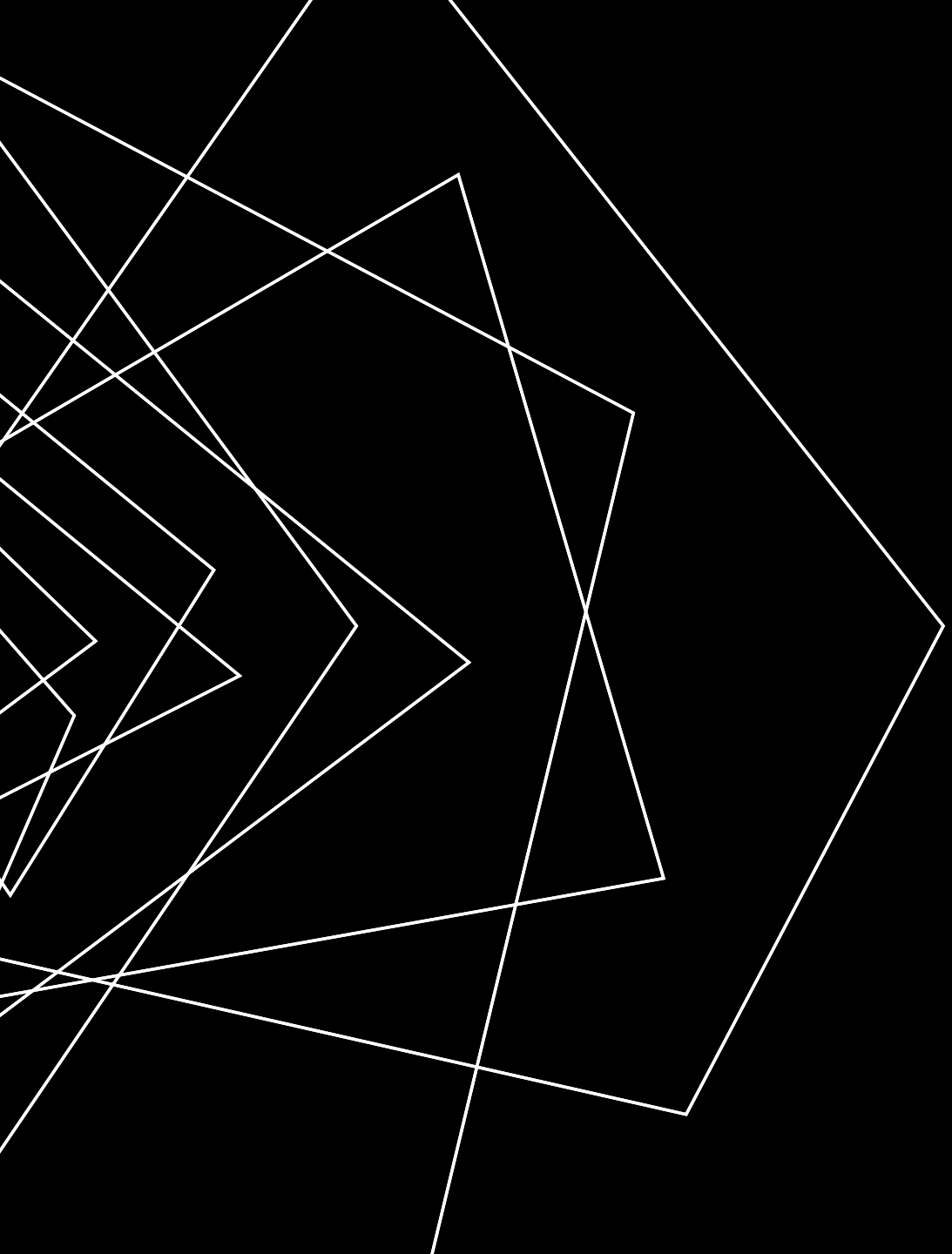
Model	Train Cross-Val Score	Test Accuracy Score	Train Roc_Auc	Test Roc_Auc	Train f1 Score	Test f1 Score
LogisticRegression_default	0.620221	0.634429	0.773574	0.766667	0.599001	0.595560
LogisticRegression_tuned	0.621674	0.634429	0.773620	0.766910	0.599992	0.595560
LDA_default	0.630865	0.605416	0.772108	0.767393	0.606125	0.576380
LDA_tuned	0.630865	0.605416	0.772108	0.767393	0.606125	0.576380
KNN	0.612477	0.622824	0.813885	0.741279	0.646744	0.596375
SVM_default	0.619262	0.620890	0.774671	0.747657	0.610126	0.581319
SVM_tuned	0.621681	0.586074	0.757978	0.736769	0.570227	0.534741
RandomFroest_default	0.607643	0.626692	0.821413	0.765531	0.644958	0.595439
RandomForest_tuned	0.634254	0.597679	0.798901	0.765361	0.616658	0.560172
XGBoost_default	0.611030	0.618956	0.820456	0.763966	0.644022	0.590477
XGBoost_tuned	0.633768	0.599613	0.779969	0.762454	0.613530	0.569692
LogisticRegression_SMOT	0.553827	0.547389	0.753664	0.752051	0.566318	0.560876
RandomForest_SMOT	0.566821	0.529981	0.783039	0.749402	0.592775	0.542708

PERFORMANCE COMPARISON ACROSS ALL MODELS



SUMMARY

- **Logistic Regression:**
  - Best overall performance with balanced metrics
- **Random Forest & XGBoost:**
  - Strong in Train Roc_AUC and F1 Scores, but overfitting observed
- **Tuning Impact:**
  - Minimal effect on linear models like Logistic Regression and LDA
  - Significant improvements in training performance for tree-based models, but overfitting limits test gains.



THANK YOU

LI WU

METRO COLLEGE OF TECHNOLOGY

INSTRUCTOR: AMIT KUKREJA

NOV. 20, 2024