Stay or Go

Job Change Predictions of Data Scientists

By Melvin Garcia
July 14, 2021
Flatiron Cohort - onl01-dtsc-pt-011121

Outline

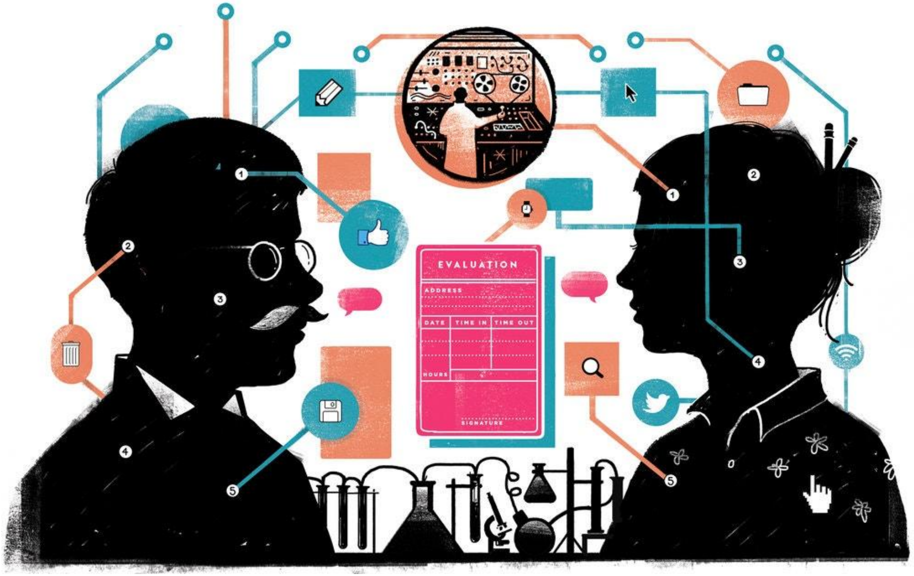
- Business Problem
- Data
- Methods
- Results
- Conclusions
- Next Steps

Business Problem

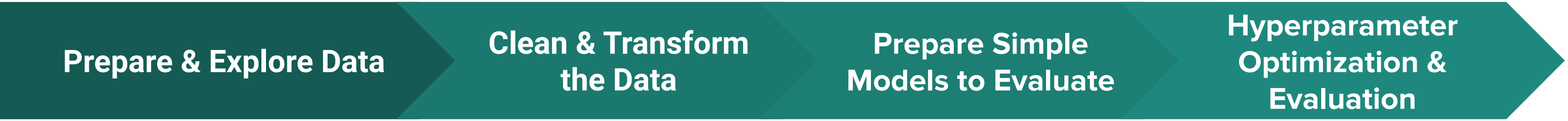
- Data Science company is looking to understand the factors that lead an employee to look for a new role or not
- The same company is conducting data science training as a service for other companies
- The objective of being able to predict if an employee will look for a new job is to help reduce the cost, time, and quality of training

Data

- Data comes variety of human resources departments containing personal information about employees participating in DS training
  - City Development Index
  - Training Hours Completed
  - Years of Experience
  - Company Size
- Dataset is imbalanced
- Most features are categorical (nominal, ordinal, binary)
- ~30% missing data contained in 2-3 features



Methods



Prepare & Explore Data

1. Understand the data types, distributions, and amount of missing data
2. Develop data strategy for encoding categorical data, and setting up transformer pipeline

Clean & Transform the Data

1. Perform appropriate transformations towards numeric features
2. Test different methods of missing value imputation
3. Prep categorical encoding pipeline

Prepare Simple Models to Evaluate

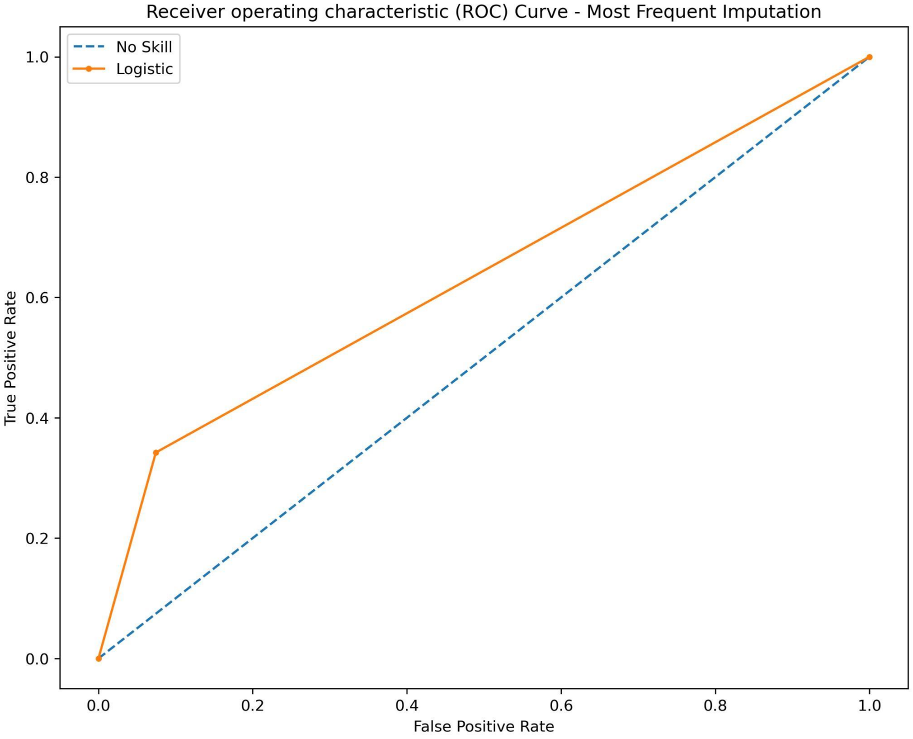
1. Prepare pipeline to test against a series of simple models
2. Evaluate the simple models and hypertune the parameters for the best performing simple model
3. Feature Importances

Hyperparameter Optimization & Evaluation

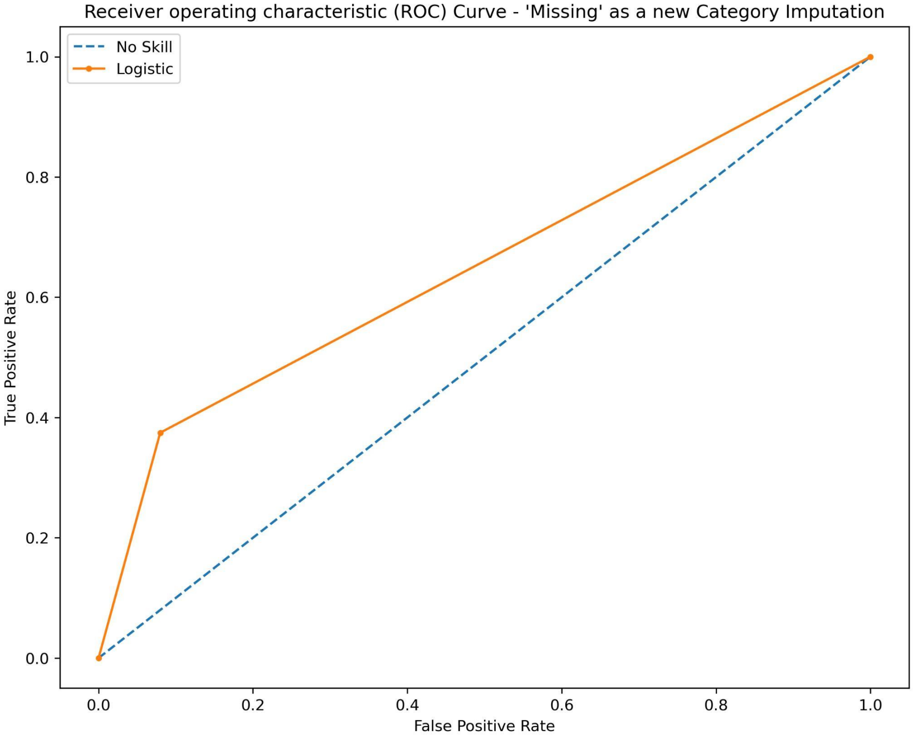
1. Take 2 of the best performing simple models
2. Perform GridSearch hyperparameter optimization
3. Evaluate any classification performance

Missing Value Imputation - LR Simple Models

Logistic: ROC AUC=0.634

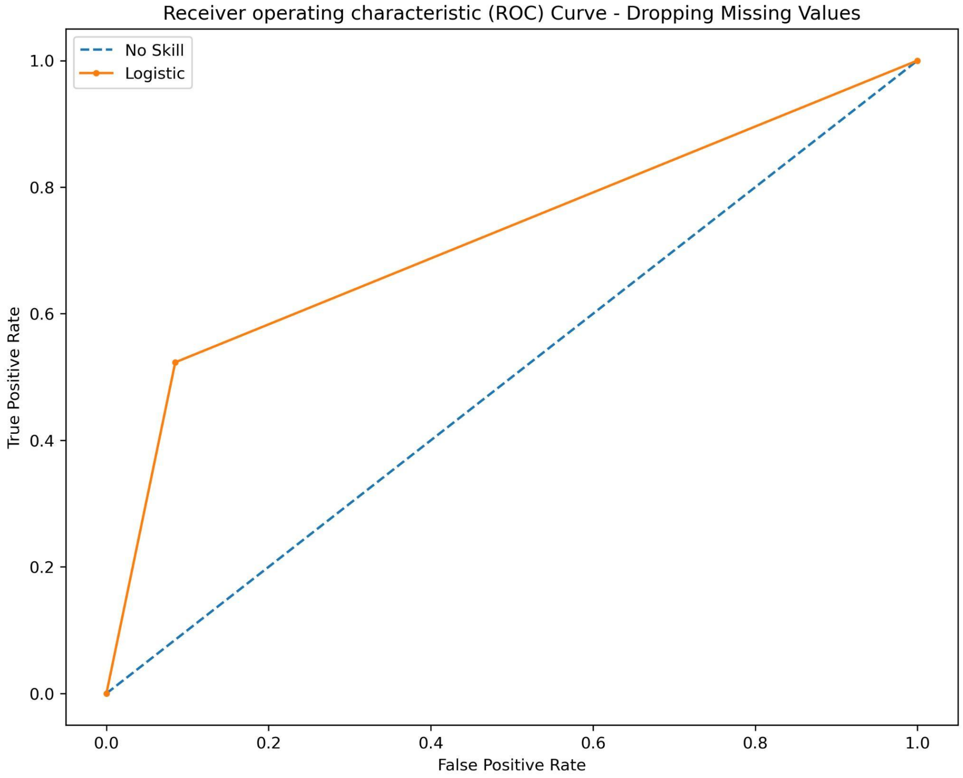


Logistic: ROC AUC=0.647

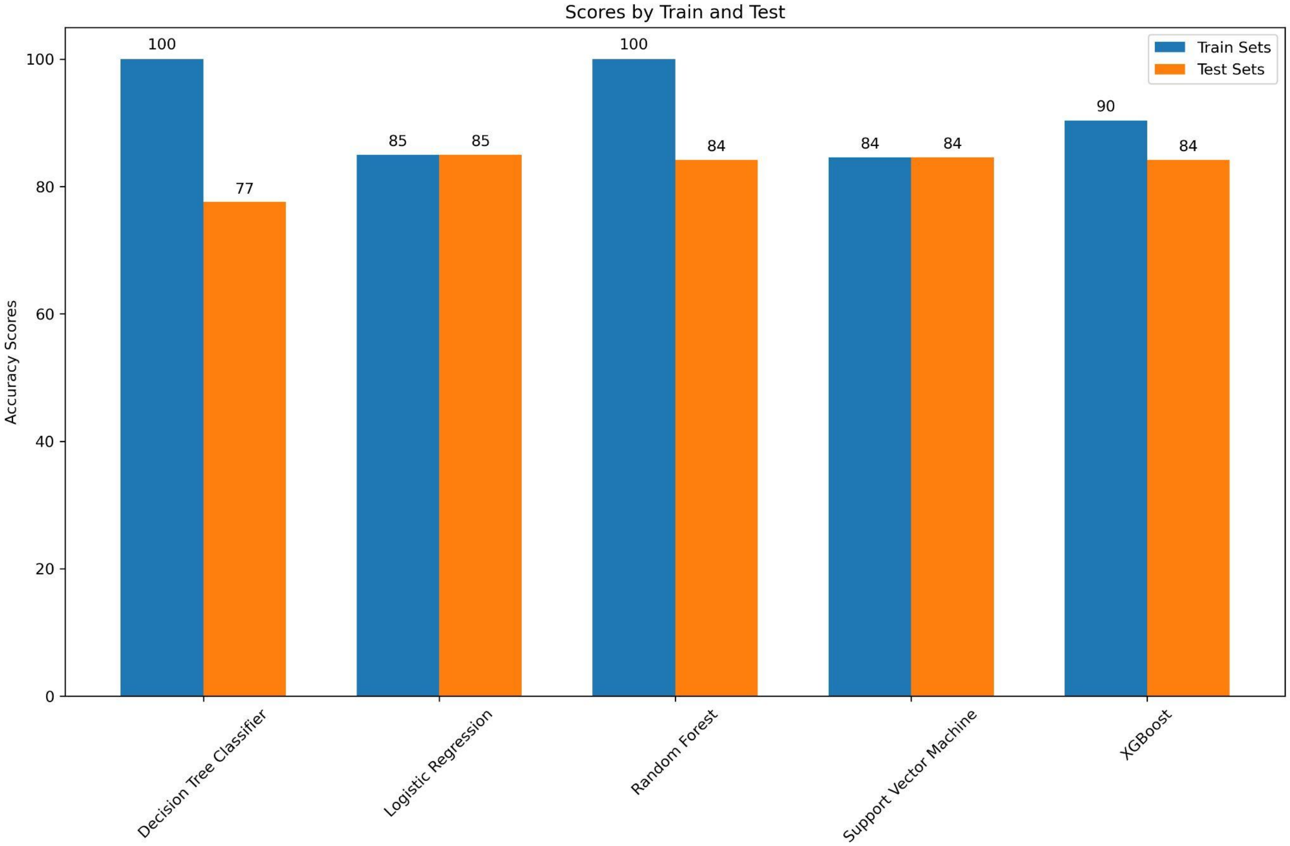


Missing Value Imputation - LR Simple Models

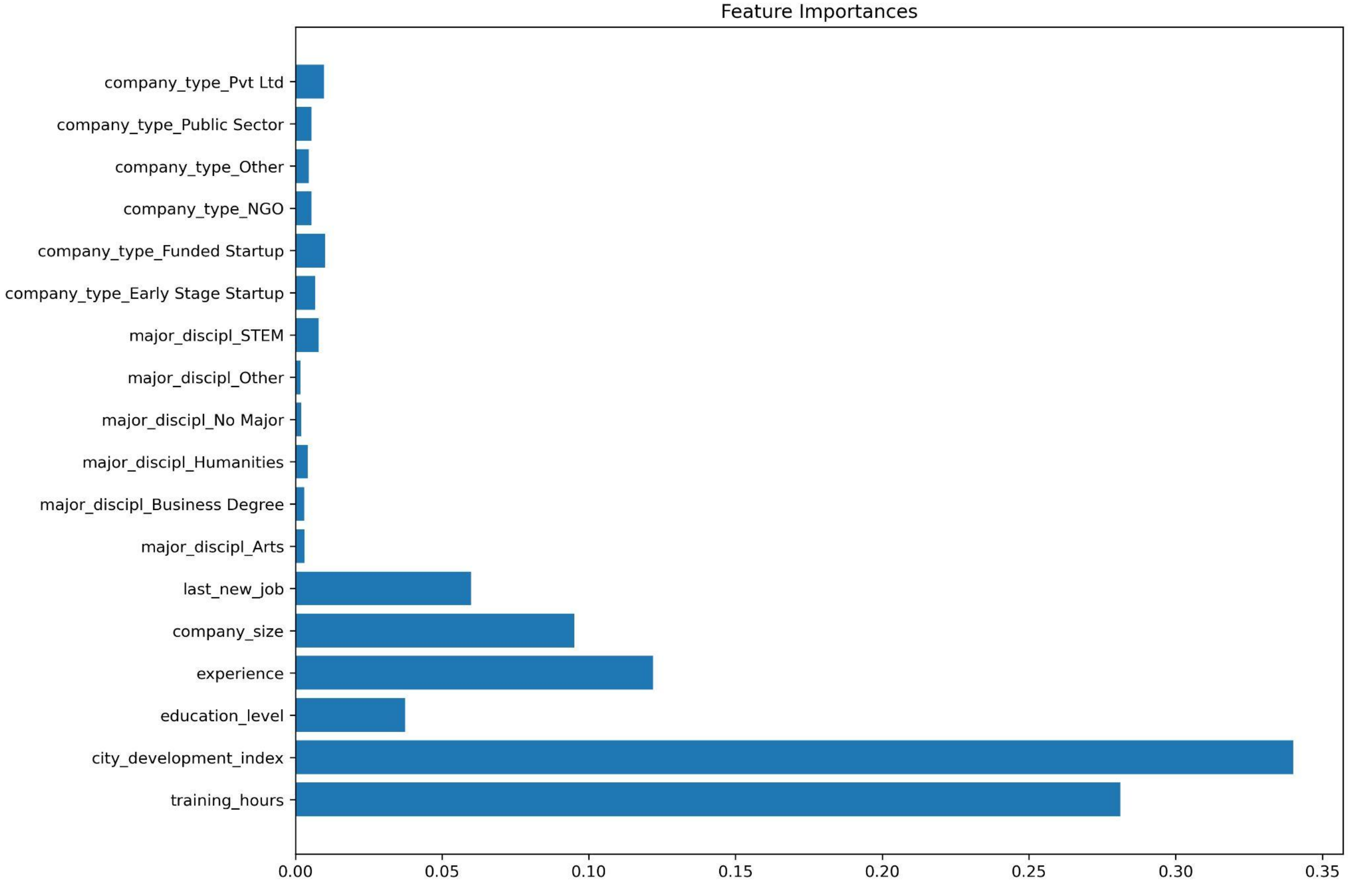
Logistic: ROC AUC=0.719



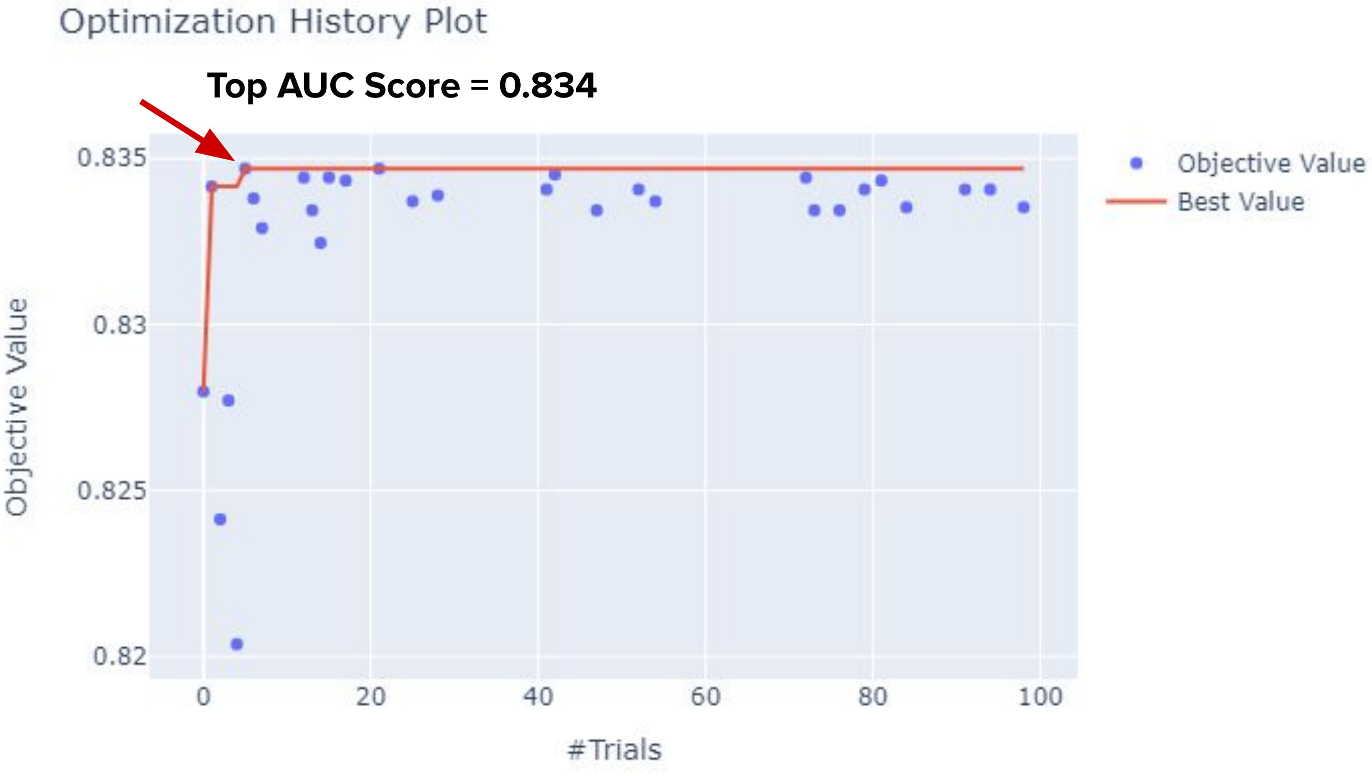
Simple Model Evaluation



Feature Importances



XGBoost Hyperparameter Tuning Results

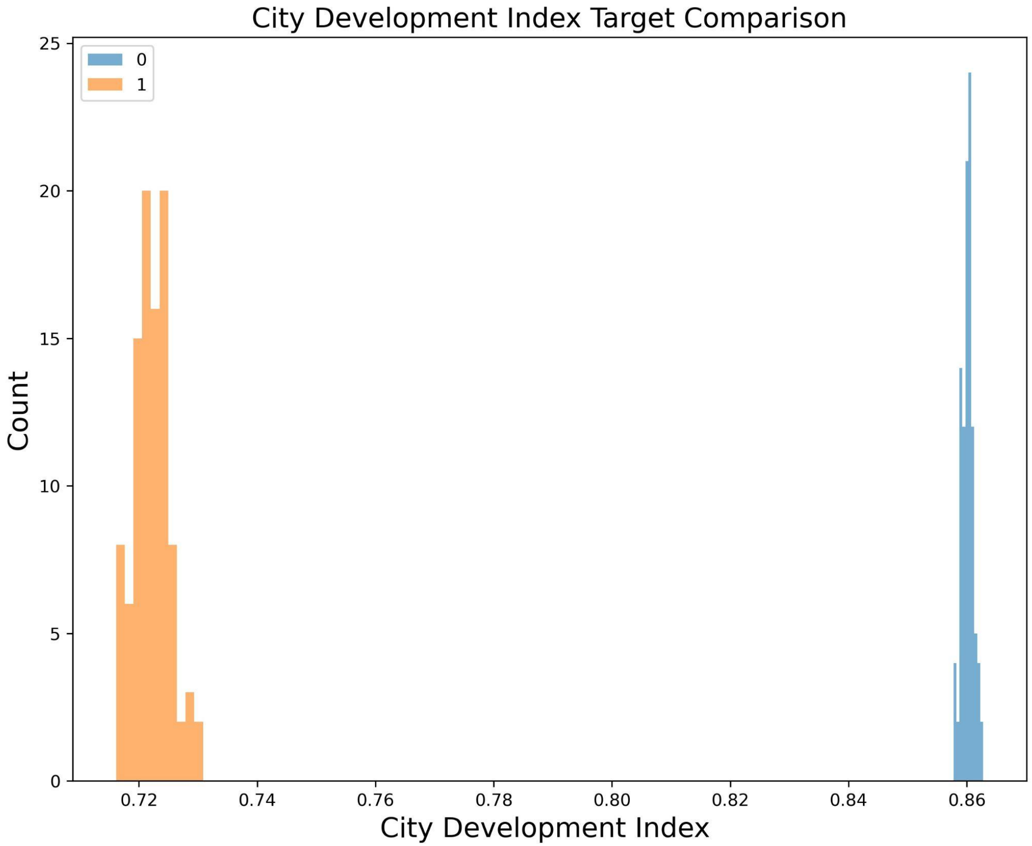


Top AUC Score = 0.834

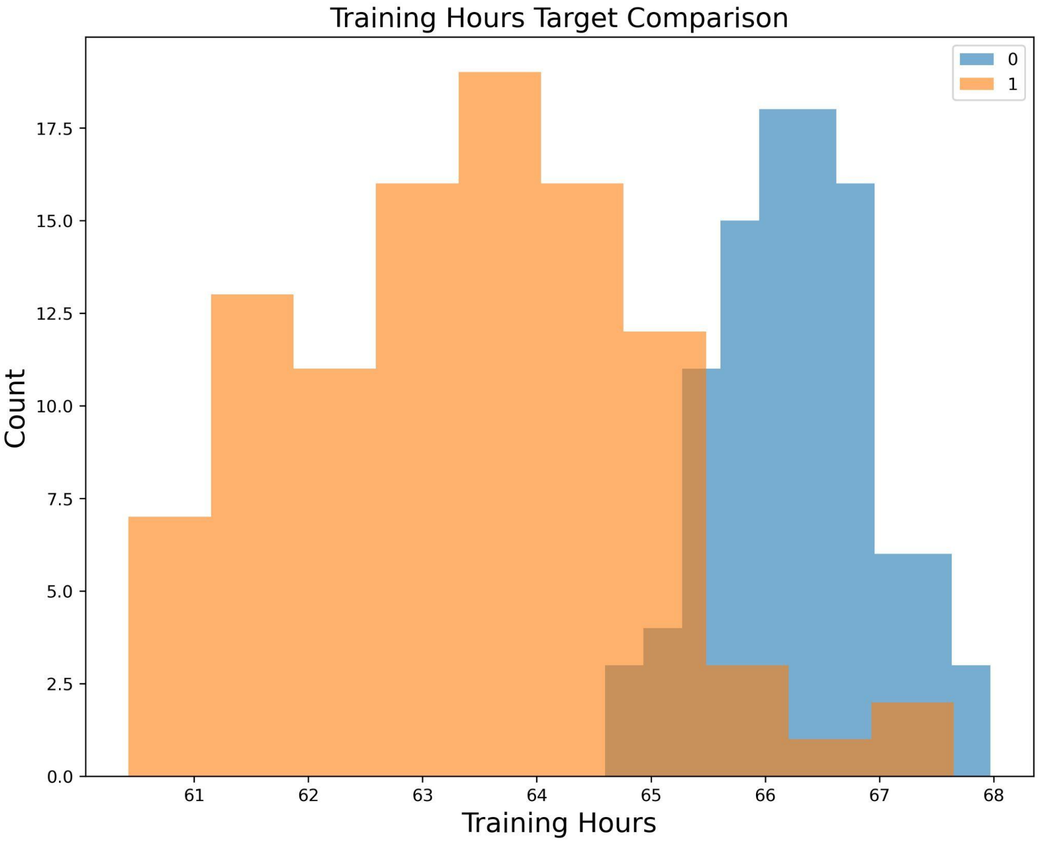
XGBoost Hyperparameter Tuning Results

Class	Precision	Recall	F1-Score	Support
Not Looking for a New Job (0)	0.88	0.93	0.90	2293
Looking for a New Job (1)	0.57	0.43	0.49	503
Accuracy			0.84	2795 (Total)

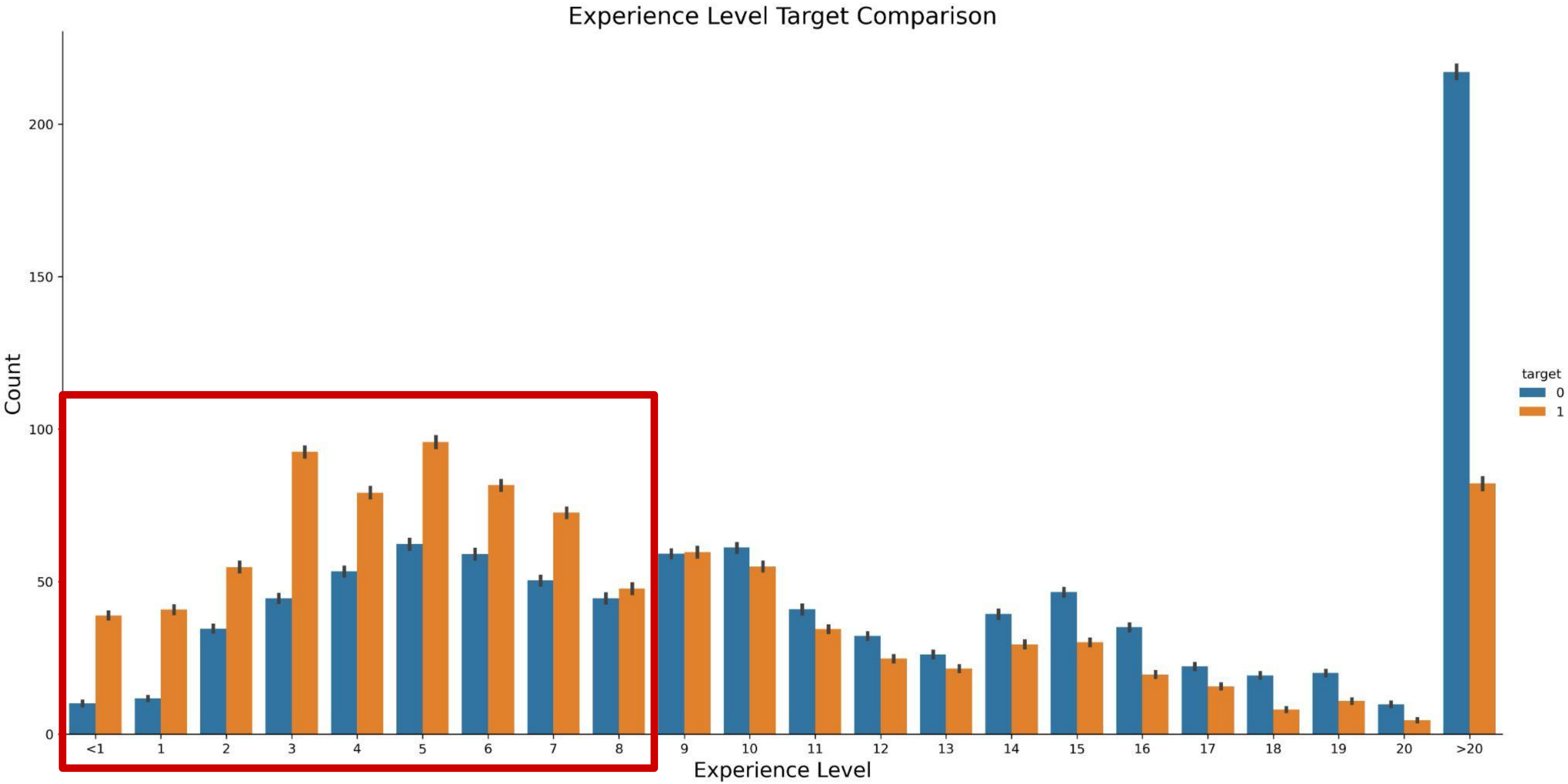
Feature Importance Deep Dive - City Development Index



Feature Importance Deep Dive - Training Hours



Feature Importance Deep Dive - Years of Experience



Conclusions

- The top three features that are observed as a factor in an employees’ decision to look for a new job are:
  - City Development Index
  - Training Hours Completed
  - Years of Experience
- The imbalance within our target class is prevalent, resulting in poor recall metrics
- Recall is a metric to optimize given the company’s objective to reduce cost and lost time for employees looking for a new role

Next Steps

- Continue experimenting with other methods of missing data imputation
- With a collective effort from the participating companies, advocate for higher data quality, especially around missing data
- Similarly, collect more data on employees who are indeed looking for a new role to help counter the imbalance of the dataset

Thank You!

Email: garciamelvin4@gmail.com
GitHub: @melvyg
LinkedIn: linkedin.com/in/melvinmgarcia/