Classifying Data Scientists Who Desire a Job Change

Juliet Womack

STAT 8240: Classification Project

November 28, 2021

Introduction

Problem Statement

With the growing number of data and analytic academic programs, bootcamps, and online resources (i.e., YouTube, Udemy, Coursea, Data Camp, etc.), the field of data science is booming. According to the U.S. Bureau of Labor Statistics (2021), data-related jobs are expected to grow by 22% in the next decade. With growth in the field of data and data science comes new, potentially lucrative opportunities for new and experienced data scientists. Although this is good for job seekers, for companies with top data talent already in their possession, they may be concerned that their top talent will leave for a new opportunity.

To ease concern, this classification project aimed to classify whether a data scientist was planning to stay at their job or was seeking a job change. A model that can accurately classify the intentions of a data scientist could help managers, team leaders, and or human resources monitor talent that is at-risk for leaving their company. Alternatively, it can help company recruiters and hiring managers stay on top of hiring new data scientists to fill in the roles of resigned data scientists prior to a data scientist's resignation.

Data Source

With HR data being subject to privacy and confidentiality laws, the present data are synthetic with practical fields and values (Möbius, 2020). The key response variable was Target, which was a binary feature, either 0, not looking for a job change, or 1, looking for a job change.

The data was pre-split into training and testing datasets by the dataset creator. The training data had 19,158 rows and the test data had 2,129 rows (approximately 90:10 split). The training data was further split into training and validation sets with the training dataset having 17,242 rows and the validation dataset having 1,916 rows (90:10 split of the original training dataset). The target variable was stratified during the train-test-split process to ensure an approximate ratio of the binary classes among the new training and validation datasets.

There were 12 features, not including the response variable, used in the subsequent data cleaning, feature engineering, and modeling process (see Table 1). Additionally, Enrollee ID was not included in the analysis or model building process because it is a unique identifier and cannot not provide meaningful insight to the classification project.

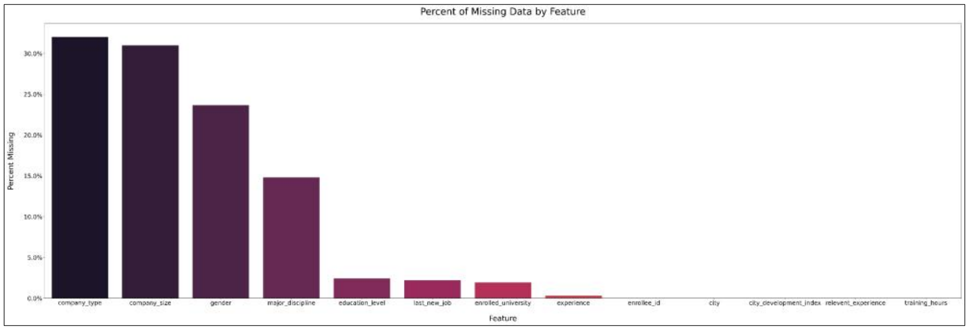
Table 1: Features from Data Scientist Job Change Dataset

Feature Name	Feature Description
Enrollee ID	Identification number for the employee. Not used in analysis.
City	The city code where the employee's employment is at.
City Development Index	Development index of the city the employee's employment is at. Values range from 0.45 to 0.95.
Gender	The gender of the employee. Either Male, Female, or Other.
Relevant Experience	Whether the employee has relevant experience or no relevant experience.
Enrolled University	Whether the employee is currently enrolled in university or not. Either no enrollment, full-time course, or part-time course.
Education Level	The employee's current education level. Either Graduate, Masters, Primary School, High School, or Ph.D.
Major Discipline	The major the employee pursued during university. Either STEM, Humanities, Arts, Other, No Major, or Business Degree.
Experience	Total amount of experience (in years) an employee has had in data science. Values are between less than 1 year and greater than 20 years of experience.
Company Size	The size of the company the employee is currently working for. Values are between less than 10 to 10,000+ employees.
Company Type	The type of company the employee is working for. Either Pvt Ltd, Funded Startup, Early State Startup, NCO, Other, or Public Sector.
Last New Job	How many years it has been since the employee work in their previous job. Either never to greater than 4 years since last job.
Training Hours	How many training hours an employee has received. Values range between 1 and 334 hours.
Target	Target variable for classification project. Either 0 or 1 for not looking for a job change or looking for a job change, respectively.

Concerns with the data are:

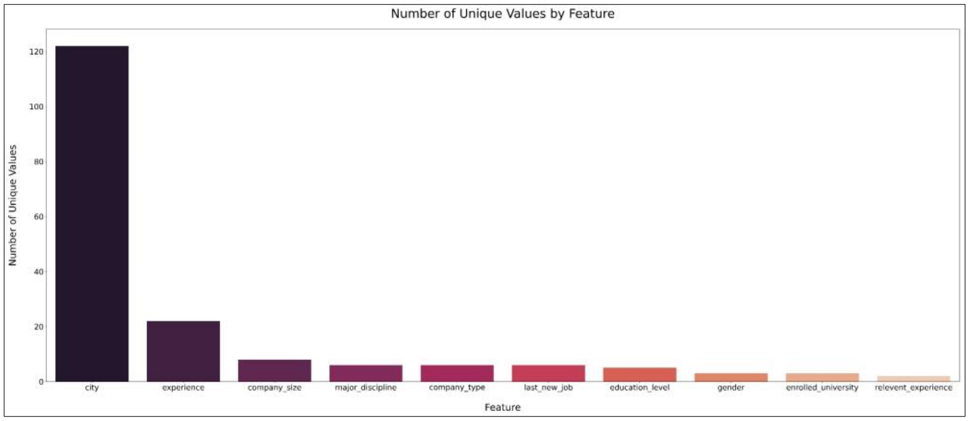
- Missing values in 8 out of the 12 features (see Figure 1).

Figure 1: Percent of Missing Data by Feature in Training Data



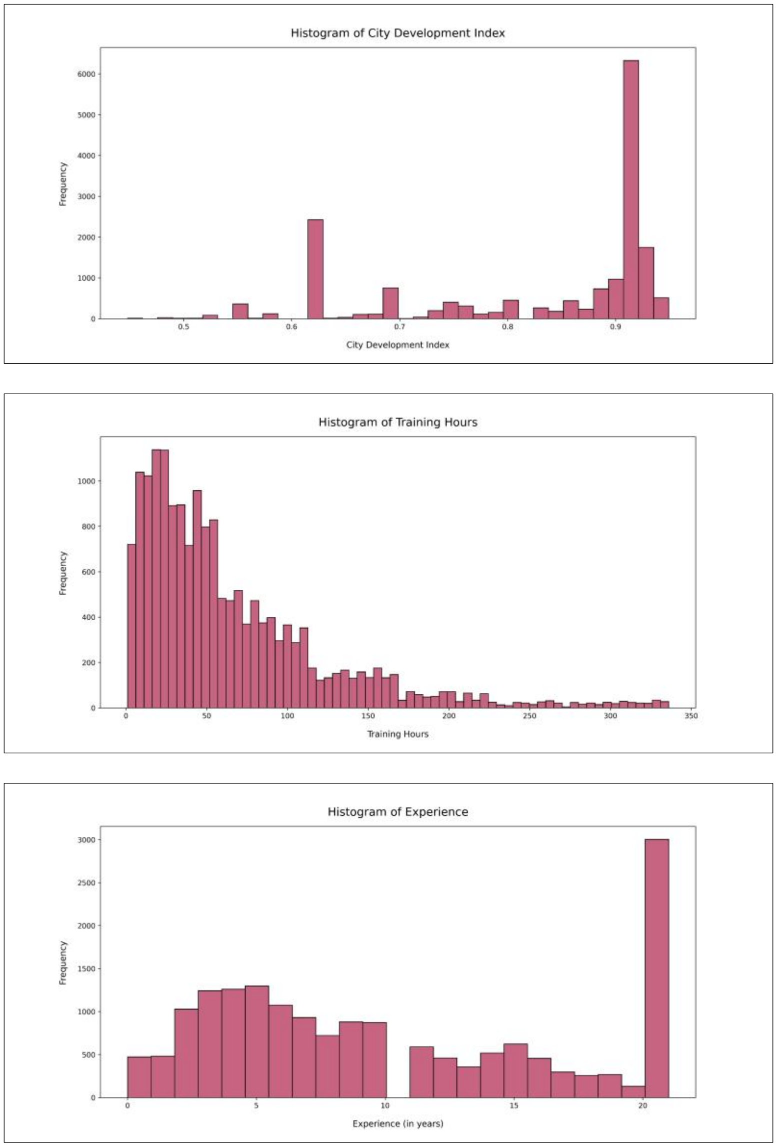
- High cardinality some of the categorical features (see Figure 2).

Figure 2: Number of Unique Values by Feature in Training Data



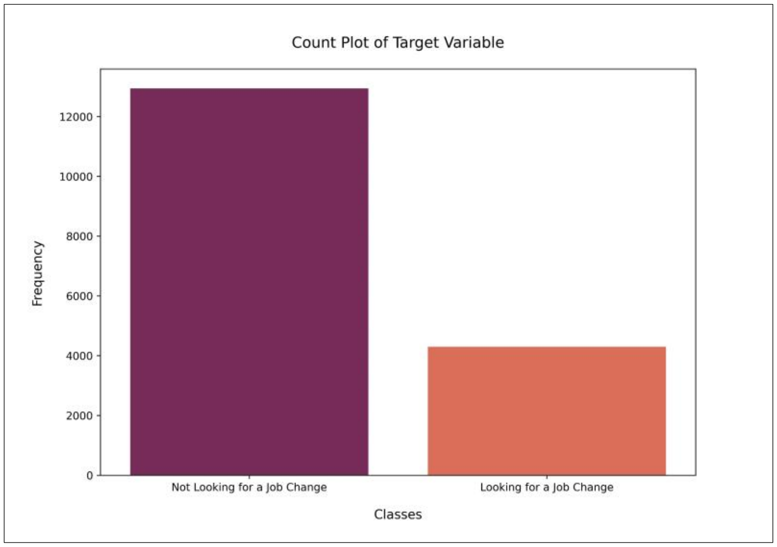
- Potential outliers in three of the numerical features (see Figure 3 for histograms of the numerical features).

Figure 3: Histograms of City Development Index, Experience, and Training Hours



- Imbalance of the target variable:
  - There were 12,943 instances of Class 0, not looking for a job change, and 4,299 instances of Class 1, looking for a job change, in the training dataset (approximately 75:25 split see Figure 4).

Figure 4: Count Plot of Target Variable



Data Preparation

Addressing Data Concerns

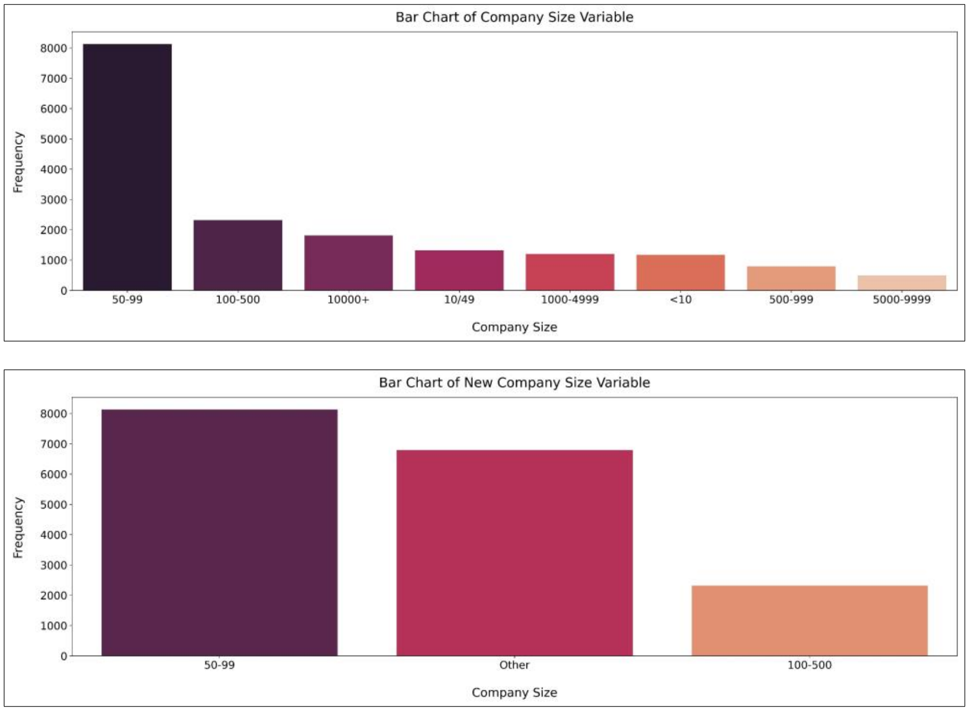
Missing Data

All features with missing data were categorical and each of these features had less than 35% of data missing. Thus, it was appropriate to fill the missing values of each feature with its respective most frequent value.

High Cardinality

For the features City, Major Discipline, and Company Size, the two categories with the highest frequencies for each feature were kept as their own category, and all other categories were combined into an "Other" category (see Figure 5).

Figure 5: Example of Reducing Variable with High Cardinality (Company Size Feature Old vs. New Variable)



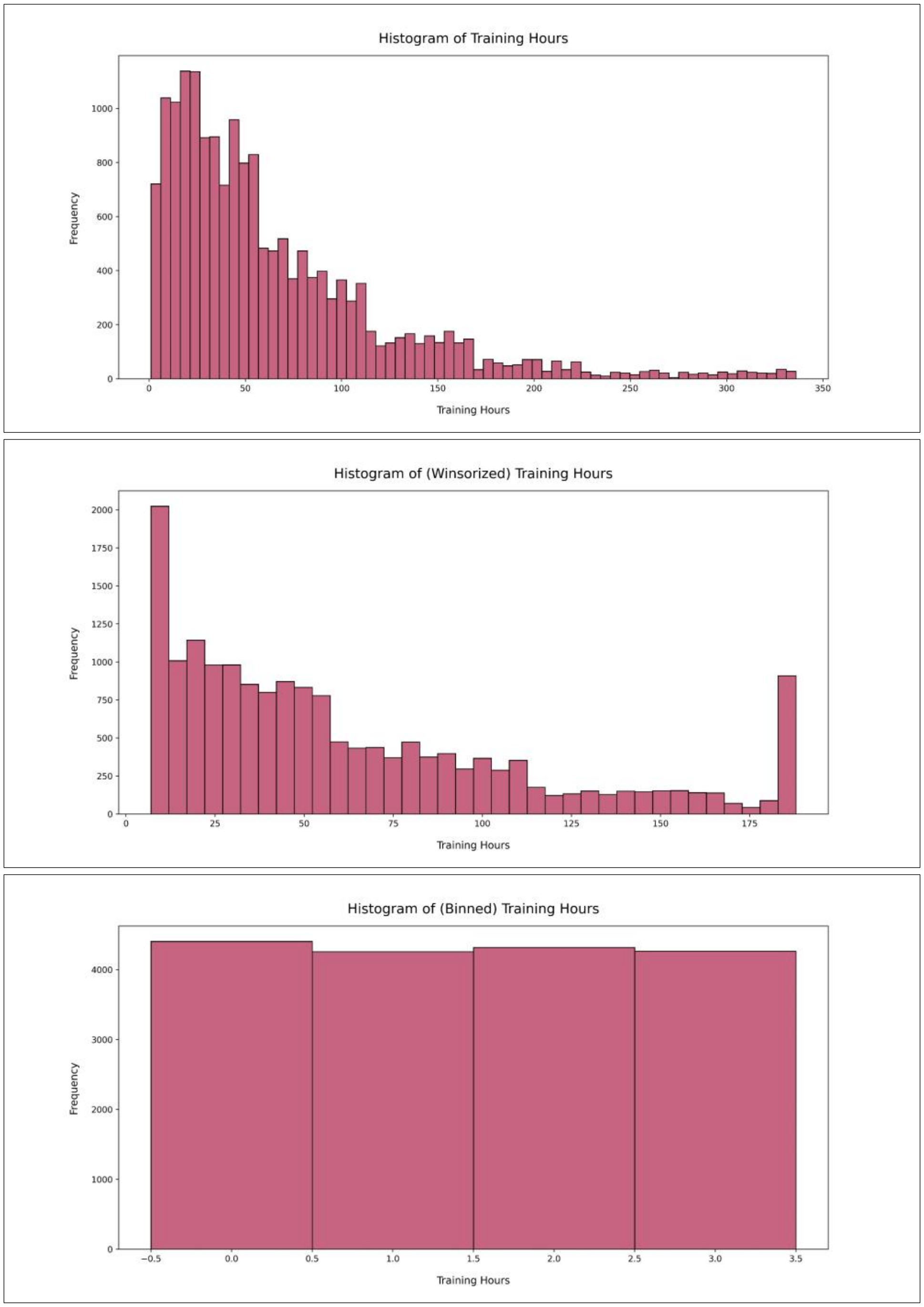
For the Education Level variable, the categories, Masters and Graduate, were combined into a single category, Masters, and then the variable was ordinally encoded. The variables Experience and Last New Job were converted from categorical to numerical.

Outliers

To handle potential outliers for the variables City Development Index, Experience, and Training Hours, a winsorized (or capped) and a binned variant were created for each. Variables were winsorized at the 5th and 95th percentile value and the variables were binned into four (approximately) equal sized bins (see Figure 6).

Figure 6: Example of Winsorizing and Binning a Variable

(Training Hours Original vs. Winsorized vs. Binned Variant)



Histogram of Training Hours

Histogram of (Winsorized) Training Hours

Histogram of (Binned) Training Hours

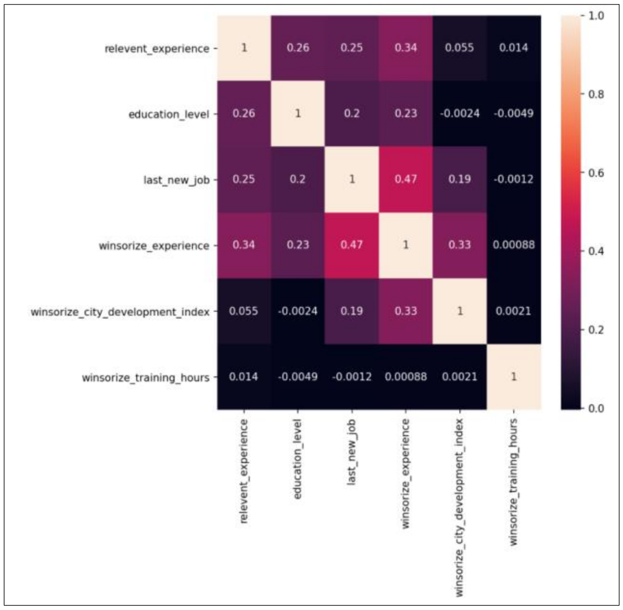
Imbalanced Target Variable

To compensate for the class imbalance, the minority class, looking for a job change, was oversampled, in the training dataset only, using the Synthetic Minority Oversampling Technique, SMOTE. SMOTE rebalanced the classes to a 50/50 split between the two classes (12,943 instances of Class 0 and 12,943 instances of Class 1 in the training dataset).

Feature Engineering

- A full list of original and feature engineered variables can be found in the Appendix in Table 1A.
- The heatmap, as shown below in Figure 7, was investigated for pairs of variables with a correlation greater than 0.30. Three new interaction variables were created:
    - Experience (winsorized) * Relevant Experience
    - Experience (winsorized) * Last New Job
    - Experience (winsorized) * City Development Index (winsorized)

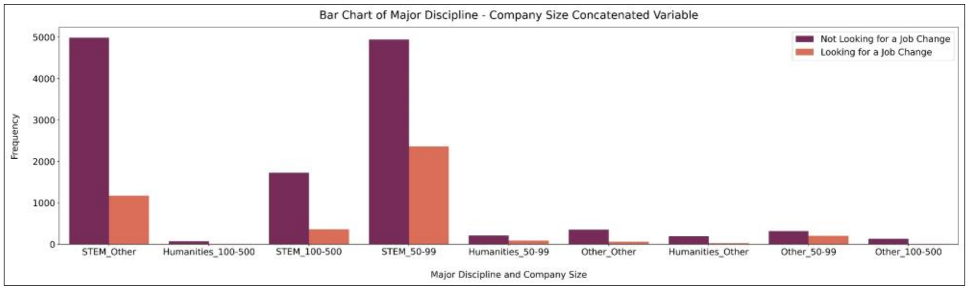
Figure 7: Heatmap of Numerical Variables in Training Dataset



- A report from LinkedIn in 2019 stated that employees were more likely to stay at their current company if given more training or opportunities to build skills in their field (Hess, 2019). Driven from that finding, five new variables were engineered using Training Hours:
    - Percentile of Training Hours by Employee
    - Training Hours Rank by Employee
    - Training Hours Categorized
        - Below average (Training Hours less than 63)
        - Average (Training Hours equal to 63)
        - Above average (Training Hours greater than 63)
    - Interactions:
        - Training Hours * Experience
        - Training Hours * City Development Index
- Indeed (2021) highlighted one of the main reasons employees leave their jobs was the desire for new, inspiring work. Thus, it may be probable that employees with more or less experience may be more likely to leave their current job due to feelings of stagnation. Based on this reasoning, two new features were engineered using Experience:
    - Experience Categorized:
        - Below Average Experience (less than 10 years of experience)
        - Average Experience (10 years of experience)
        - Above Average Experience (above 10 years of experience).
    - Percentile of Experience (in years) by Employee

- With there being several nominal features (i.e., City, Gender, Enrolled University, Major Discipline, Company Size, and Company Type), each nominal feature was paired with one another to create 15 concatenated features (see Figure 8):
    - It could be possible that a combination of nominal features is better at classify if a data scientist is looking or not looking for a job change rather than a individual feature.

Figure 8: Example of New Concatenated Variable (Major Discipline and Company Size) Stratified by the Target Variable



Model Development

Methods Used

Models were built using the following seven classification methodologies on the training dataset with and without Principal Component Analysis (PCA) applied:

- Logistic Regression (LR)
    - Coefficients can be interpreted.
    - Ideal for a binary target variable.
    - Good baseline to compare more sophisticated algorithms to.
- Decision Tree (DT)

    - Algorithm can be visualized and easy to interpret and understand the algorithm’s process to obtain model.
    - Algorithm performs feature selection by itself.
- Random Forest (RF)
    - Alternative to DT that prevents overfitting and may improve the accuracy from DT.
- Support Vector Machine (SVM)
    - May not overfit compared to other machine learning algorithms, but was the most time-consuming algorithm to implement.
- K-Nearest Neighbors (KNN)
    - Easy to understand compared to the XGBoost, SVM, or Multilayer Perceptron algorithms.
    - Adapts to new data quickly and is efficient during the training process.
- XGBoost (XBG)
    - Fast and sophisticated algorithm that can be used for multiple different machine learning challenges.
    - Implemented in case the other methods were unable to perform well.
- Multilayer Perceptron (MLP)
    - Works well with large datasets and ideal for classification problems. However, was challenging and time-consuming to find the optimal parameters to optimize model’s performance.

Model Assessment

Model Assessment Metrics of Choice

Area Under the Curve (AUC) and Recall were the two metrics that determined the winning model.

- A model with a high AUC is better at deciphering which class an individual data scientist fell under based on their characteristics.
- A model with a high Recall score is better at identifying data scientists who are truly seeking a job change.
    - It is critical for this model to identify these individuals because a loss of these data scientists could reduce human capital for the company and result in more time, money, and resources spent recruiting, hiring, and training new data scientists.

Model Preparation, Variable Selection, and Principal Component Analysis

Prior to model implementation, the data were:

- Dummy encoded for the categorical features
- Scaled using Standard Scaler.
- Variable reduction was applied using Random Forest feature importance. The number of variables included in the model were reduced from 73 to seven (see Figure 9).
    - For PCA implementation, the number of components selected was seven.

- For variable selection using Random Forest, the top seven variables were chosen because this number of features that maximized the AUC while still producing a relatively simple model (see Figure 10).

Figure 9: Top 7 Features from Random Forest Feature Importances

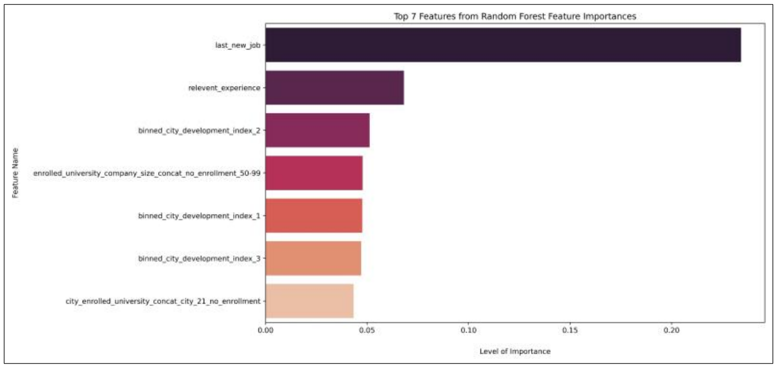
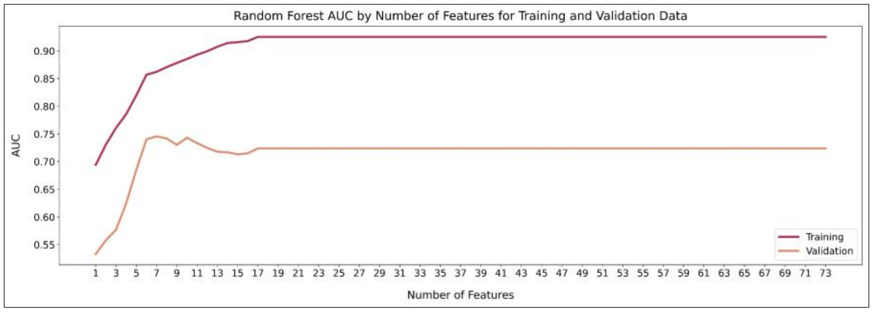


Figure 10: Random Forest Classifier AUC by Number of Features Included in Training and Validation



Base Models

All seven methodologies were implemented with no additional parameter tuning on the training dataset with no PCA applied. The model with the highest AUC and Recall score, the MLP Classifier Model, was subject to further parameter tuning. Figure 11 and Figure 12 shows the AUC and Recall scores of all seven base models, respectively. The winning model, the Base MLP Classifier model, had an AUC of 0.747 and a Recall Score of 0.656.

Figure 11: AUC Scores for the 7 Base Models

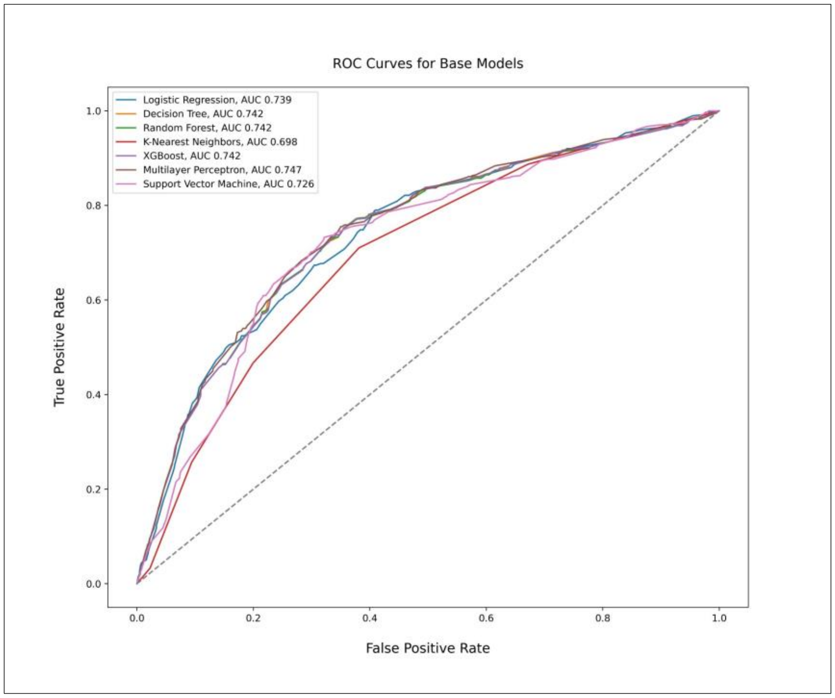
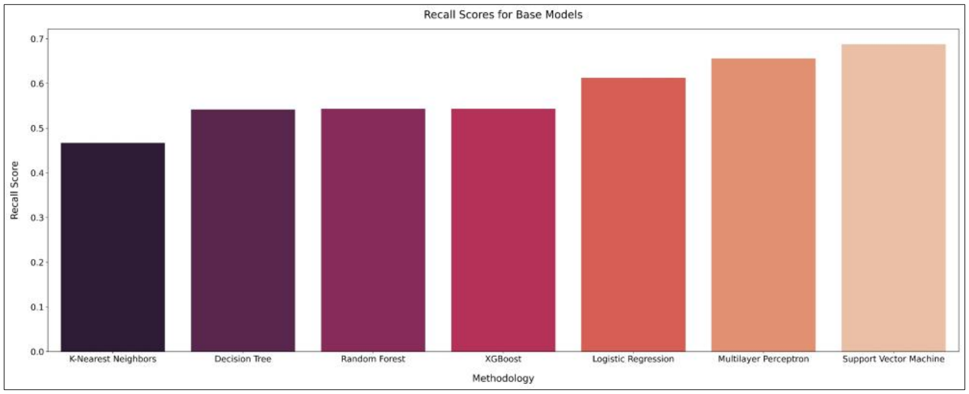


Figure 12: Recall Scores for the 7 Base Models



PCA Models

Using PCA with seven components, all seven methodologies were implemented with no additional parameter tuning. Similar to the base models, the model with the highest AUC and Recall score, the PCA MLP Classifier, was selected as the winning model and would be compared to the base MLP model and parameter tuned MLP models. Figure 13 and Figure 14 shows the AUC and Recall scores of all seven PCA models, respectively. The PCA MLP Classifier Model had an AUC of 0.757 and a Recall Score of 0.712.

Figure 13: AUC Scores for the 7 PCA Models

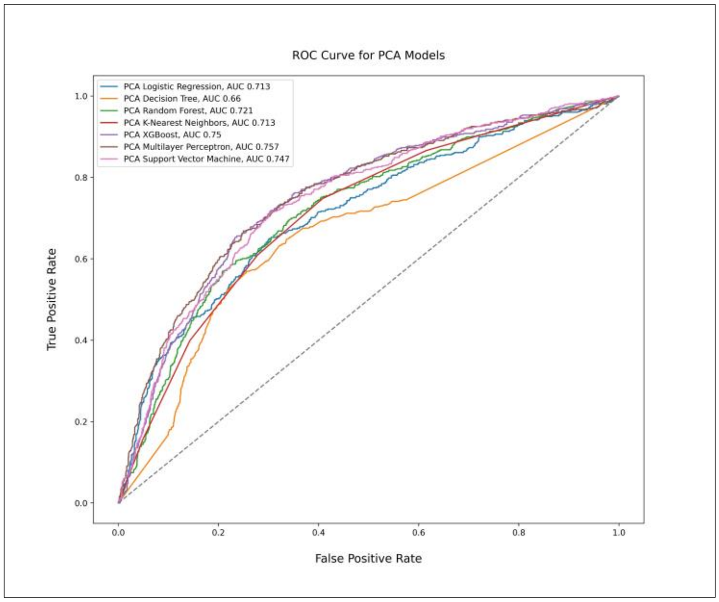
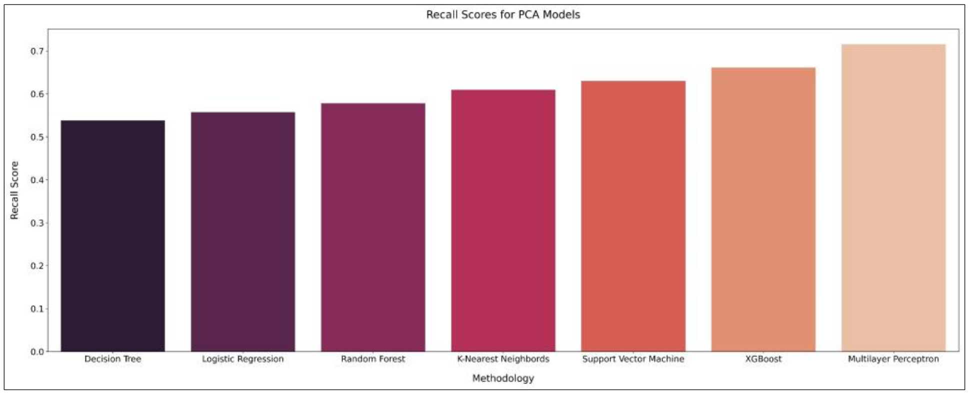


Figure 14: Recall Scores for the 7 PCA Models



MLP Models and Winning Model

The base and PCA MLP models had the best AUC and Recall scores. To, hopefully, improve model performance on the base model, Random Grid Search and Grid Search were implemented to find optimal parameters for the Base MLP classifier. The best parameters found were fitted on the full training dataset (see Appendix Table 2A for best parameters from Random Grid Search and Grid Search). Figure 15 and Figure 16 shows the AUC and Recall scores of the four MLP classifier models.

Both implementations of Grid Search were not successful at improving the Base MLP Classifier Model performance. Out of the four MLP models, based on AUC and Recall, the winning model was the PCA MLP classifier model.

The PCA model does the best at distinguishing between the two classes and identifying data scientists who are truly interested in a job change. The AUC and Recall Scores for all four MLP models are in Table 2 and the Confusion Matrix for the testing dataset for the winning model is in Figure 17.

Figure 15: AUC for the 4 MLP Classifier Models

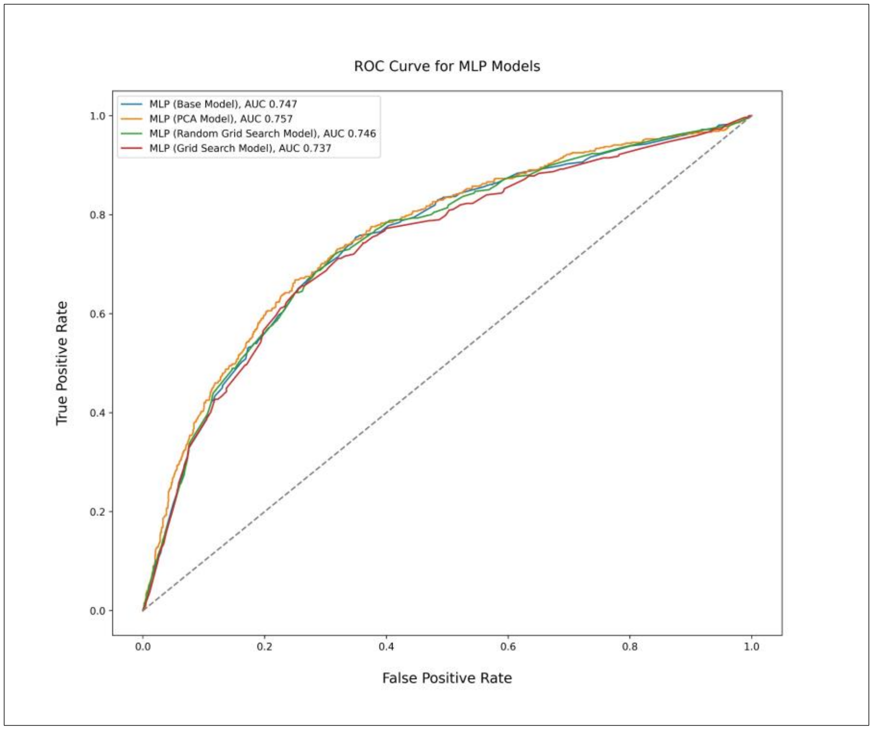


Figure 16: Recall Scores for the 4 MLP Classifier Models

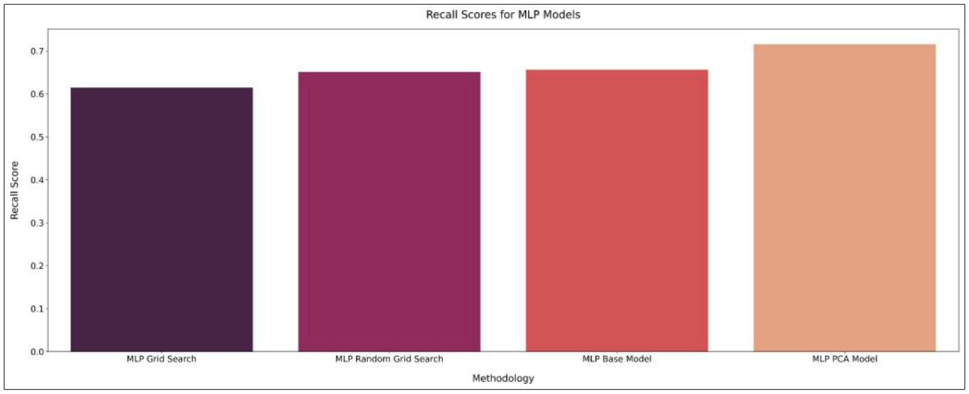
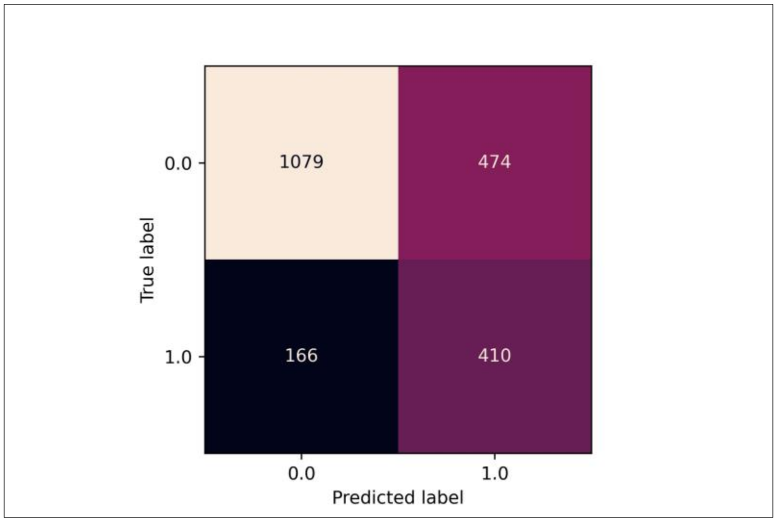


Table 2: AUC and Recall Scores for the 4 MLP Classifier Models

(Winning Model: PCA MLP Classifier Model)

Model	AUC	Recall
Base	0.747	0.656
Random Grid Search	0.746	0.651
Grid Search	0.737	0.615
PCA	0.757	0.712

Figure 17: Confusion Matrix for the Winning PCA MLP Classifier Model (Testing Data)



Conclusion

Model Result Implications

The winning model can help identify data scientists who are wanting to change jobs. However, the winning model tends to incorrectly flag a large number of data scientists who are not wanting to change jobs.

If there is not a high cost to misclassifying a data scientist, these misclassified data scientists could be subject to closer monitoring and intervention by HR or management resulting in a decrease in the number of data scientists desiring to change jobs. However, if there is a high cost to misclassifying data scientists, then the model would need further improvements to lower the false positive rate.

This model could be run on a quarterly or bi-yearly basis by HR to identify data scientists at-risk for leaving the company. Depending on the quality of the data scientist(s) wanting to change jobs, efforts could be made to improve the job satisfaction of those data scientist(s) (e.g., increase in training, raise in salary, increase in benefits, etc.) or HR could prepare to recruit new data scientist(s) prior to resignation of these employees. Overall ensuring the data scientist team or department for the company has quality talent year-round.

References

Bureau of Labor Statistics, U.S. Department of Labor. (2021, September 8). Occupational Outlook Handbook: Computer and Information Research Scientists. *U.S. Bureau of Labor Statistics*. https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm

Hess, A. J. (2019, February 27). LinkedIn: 94% of employees say they would stay at a company longer for this reason—and its not a raise. *CNBC*. https://www.cnbc.com/2019/02/27/94percent-of-employees-would-stay-at-a-company-for-this-one-reason.html

Indeed.com. (2021, February 24). 16 Reasons employees leave their jobs. *Indeed.* https://www.indeed.com/career-advice/career-development/reasons-employees-leave

Möbius. (2020, December 6). HR analytics: Job change of data scientists (version 1). *Kaggle*. https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

Appendix

Table 1A: Full Variable List

Variable	Original or Engineered
City	Original
Gender	Original
Relevant Experience	Original
Enrolled University	Original
Education Level	Original
Major Discipline	Original
Company Size	Original
Company Type	Original
Last New Job	Original
Experience (Winsorized)	Original
Experience (Binned)	Engineered
City Development Index (Winsorized)	Original
Binned City Development Index	Engineered
Training Hours (Winsorize)	Original
Training Hours (Binned)	Engineered
Experience (in years), Relevant Experience Interaction	Engineered
Experience (in years), Last New Job Interaction	Engineered
Training Hours Percentile	Engineered
Training Hours Rank	Engineered
Training Hours Categorized	Engineered
Training Hours, Experience Interaction	Engineered
Training Hours, City Development Index Interaction	Engineered
Experience (in years) Categorized	Engineered
Experience (in years) Percentile	Engineered
Experience City Development Index Interaction	Engineered
City, Gender Concatenated	Engineered
City, Enrolled University Concatenated	Engineered
City, Major Discipline Concatenated	Engineered
City, Company Size Concatenated	Engineered
City, Company Type Concatenated	Engineered
Gender, Enrolled University Concatenated	Engineered
Gender, Major Discipline Concatenated	Engineered
Gender, Company Size Concatenated	Engineered
Gender, Company Type Concatenated	Engineered
Enrolled University, Major Discipline Concatenated	Engineered
Enrolled University, Company Size Concatenated	Engineered
Enrolled University, Company Type Concatenated	Engineered
Major Discipline, Company Size Concatenated	Engineered
Major Discipline, Company Type Concatenated	Engineered
Company Size, Company Type Concatenated	Engineered

Table 2A: Best Parameters from Random Grid Search and Grid Search CV

Methodology	**Hidden Layer Sizes**	**Learning Rate**	**Batch Size**
Random Grid Search	(100, 100)	0.01	1000
Grid Search CV	(77, 77, 77)	0.001	50