Employability in Tech: What Gets You A Job Nowadays

Denicia Espinosa Aragon[1]

[1] University of Oregon

Author Note

This paper is a submission the EDLD 654: Machine Learning for Educational Data Science course at the University of Oregon. The data was obtained from Kaggle, an online data science community that host data set for public use. Exact datasource link here: https://www.kaggle.com/datasets/ayushtankha/70k-job-applicants-data-human-resource

Correspondence concerning this article should be addressed to Denicia Espinosa Aragon. E-mail: daragon@uoregon.edu

Employability in Tech: What Gets You A Job Nowadays

Research Problem

Jobs in tech are rising in popularity by the day while steadily decreasing in availability at the same rate Picciotto (2023). The news has continuously featured headlines of mass lay-offs at major companies such as Microsoft, Meta, Amazon, General Motor to name a few, totaling over two million tech professional layoffs in 2023 alone Jose (2023). Besides the overwhelming frequency of layoffs, tech jobs are still, again, high in popularity and further continuing to be in high-demand Picciotto (2023). Amongst the change in the job market and massive company reorganizations leading to layoffs, tech professionals must be able to keep up with a changing market, and to do so they need to understand what skills and attributes are needed to stay competitive.

On the other side of the coin, it equally important for companies to know how the industry overall is determining employability, or in other words, what features of a resume contribute the most to getting hired across major tech companies. Understanding employability patterns can address two concerns, reducing the increased labor needed in the hiring process and ensuring equitable decision-making in hiring.

Given the shift in the job market and the surge in job applications, resulting in thousands of submissions for a single job title, companies must streamline operations to enhance efficiency in the hiring process. Efficiency can be provided by extracting crucial information from resumes and assessing it against established criteria used in previous processes for selecting candidates for interviews. Further, extracting demographic data from resumes can help companies evaluate bias in their employment selection processes and make improvements.

The present project intends build and compare multiple machine learning models to effectively predict employment of programmers based on a variety of classic resume

components. Additionally, the project will investigate the paramount elements that influence employability. ## Description of Data

The present data was obtained via Kaggle, "70k+ Job Applicants Data (Human Resource)" (n.d.) and was created from StackOverlow's annual user-generated survey that includes 73462 software developers in over 180 countries. At download, the variables were standardized to ensure consistency among data points as well cleaned for null responses. Rows in the data represent an individual response from a software developer. Columns represent variables measured for each participant. A few columns were excluded from the dataset due to the lack of detail in what these variables represent or repeated complications in analyses, including, "Accessibility", "HaveWorkedWith" and "Employment". Additionally, in the original data set, the variable "Country", representing the origin of each participant, was converted into "Continent", representing the continent of origin for each participant. Creating the "Continent" variable was conducted reduce dimensionality of the data and improve generalization. Dummy variables were created for all categorical variables and the outcome measure, "Employed" was converted into a factor variable. For full description of variables see Table 1.

Description of Models

The outcome variable, Employed, is a binary variable. Therefore, the present project utilizes classification models, specifically using unregularized logistic regression, logistic regression with a ridge penalty, and bagged decision tree for classification. Utilizing this list of models offers a diverse range of linear and non-linear models, each exhibiting distinct levels of variance and bias. Employing a unregularized logistic regression assumes there is a linear relationship between predictors otherwise known as a less complex model, which ensures a low amount of bias in the model. However, due to the low amount of bias, a unregularized logistic regression model may have higher variance and the trade-off of this balance causes the model to be prone to overfitting. Deploying a ridge logistic regression

adds a penalty to the model coefficients which reduces variance in the coefficients but adds some bias to the model to strike a balance between the variance and bias. Bagged decision trees generally have low bias thanks to the complexity of the model. The hyperparameter for a ridge logistic regression is lambda, which is the penalty placed on the coefficients. For the present dataset, an array of penalty parameters were declared in the attempt to investigate what amount of penalty was optimal. The last model type, bagging, is short for bootstrap aggregating and describes the process by which multiple models of the same learning algorithm are trained with re-sampling from the same data set. After predictions are created for the multiple models, the predictions are the aggregated to have a final prediction observation for the data set. The hyperparameter for bagged models is the number of bootstrapping samples involved in the model. To evaluate model performance and select optimal model, logloss, area under the ROC (AUG), model accuracy, model precision, and model sensitivity will be compared across the three models. The optimal model will minimize logloss and maximize all other performance parameters.

Model Fit

Based on the model performance measures, the non-regularized logistic regression is the most optimal model to predict employability for software developers, but only by a really small margin. As mentioned previously, the optimal prediction model will minimize logloss, which measures the accuracy of a classification model by penalizing false classifications. The non-regularized logistic regression had a logloss of 0.4437, a few points lower than the ridge logistic regression .4597, and the Bagged Trees logistic regression 0.4597. Further, the non-regularized logistic regression maximized AUC, precision, and true negative rates. AUC also known as area under the ROC curve, measures the model's ability to differentiate between positive and negative classes. Precision is the ratio of correctly predicted positive observations total predicted positive observations. TNR or true negative rates is the measurement how many negatives the model predicts correctly. The

non-regularized logistic regression had a AUC of 0.8713, a few points higher than the ridge logistic regression 0.8706, and the Bagged Trees logistic regression 0.8548. For precision scores, non-regularized logistic regression had a score of 0.8001, again few points higher than the ridge logistic regression 0.7995, and the Bagged Trees logistic regression 0.7742. For true negative scores, non-regularized logistic regression had a score of 0.7696, again few points higher than the ridge logistic regression 0.7994, and the Bagged Trees logistic regression 0.7284. Ridge logistic regression performed the best on accuracy (ACC) 0.7847 by a few points compared to non-regularized logistic regression 0.7836 and bagged decision trees 0.7686 while bagged decision trees performed best on true predicted rate (TPR) 0.8032 compared to non-regularized logistic regression 0.7994 and ridge logistic regression 0.7994. (See Table 2 for all numbers) For all analyses, a cut-off point/threshold of .5 was implemented.

Conclusion

While non-regularized logistic regression marginally improved only a few model performance parameters, it consistently outperformed ridge logistic regression and bagged decision trees across the majority of metrics, establishing itself as the optimal model among the three. Bagged decision trees emerged as the least optimal model. In addition to revealing the model performance, it is noteworthy that the most critical parameters vary significantly across each model. For the non-regularized regression, the most important predictor by a landslide was computer skills, again measuring the umber of computer skills known by each participant. Computer skills reported a 100 on the importance scale, possibly indicating over-fitting occurring in this model. Behind computer skills featured previous salary, if a participant was a previously a developer, obtaining a PhD and living in North America to name the first five. For the ridge logistic regression, the most important predictor was obtaining a PhD also will a reported score of 100 on the importance scale. Followed behind a PhD, if a participant was a previously a developer or not, computer

skills, and living in South America (See figures 1 and 2).

On a separate note, I learned a ton from this course. My work focuses on how individuals connect with their future self and how we can encourage good decision for a better tomorrow and this course allowed me to think differently about how we investigate open-ended data and predict outcomes in psychology. I hope to use the lessons of this course to create word embeddings and computer a semantic similarity analysis on open-ended descriptions of current and future self and correlate similarity to a measure of future self connection to see if how we describe out current and future self can describe the relationship we feel with our future selves.

References

70k+ job applicants data (human resource). (n.d.). Retrieved December 6, 2023, from https://www.kaggle.com/datasets/ayushtankha/70k-job-applicants-data-human-resource

Jose, B. (2023, December 7). Tech layoffs 2023: Job cuts deepen as spotify, amazon, yahoo and others lay off staff | technology news - the indian express. Retrieved December 7, 2023, from https://indianexpress.com/article/technology/tech-news-technology/tech-layoffs-spotify-amazon-layoffs-9056846/

Picciotto, R. (2023, July 7). Tech roles are still 'the most in-demand,' says job market expert—but you need these skills to land them. CNBC. Retrieved December 6, 2023, from https://www.cnbc.com/2023/07/07/tech-jobs-are-still-the-most-in-demand-says-employment-market-expert.html

Table 1

Variable Descriptions

Variable	Description
Age	Participant's age, categorical w/ 2 options: lessthan35,greaterthan35
Gender	Participant's gender, categorical w/ 3 options: Man , Woman, NonBinary
Continent	Participant's continent of origin, categorical
ComputerSkills	Number of computer skills known by applicant, integer
EdLevel	Education level of the applicant, categorical w/ 4 options: Undergraduate, Master, PhD, Other
MainBranch	If the applicant is a professional developer, categorical w/ 2 options: Dev, NoDev
MentalHealth	Mental Health of the participant, categorical w/ 2 options, Yes, No
PreviousSalary	Previous salary of the participant, integer
YearsCode	How long the applicant has been coding, integer
YearsCodePro	How long the applicant has been coding in a professional context, integer
Employed	If the participant was hired or not, categorical, w/ 2 options: Yes, No

Table 2

Model Performance for All Models

Model	LL	AUC	ACC	TPR	TNR	PRE
Non-Regularized Logistic Regression	0.4437	0.8713	0.7836	0.7957	0.7696	0.8001
Logistic Regression with Ridge Penalty	0.4597	0.8706	0.7847	0.7994	0.7677	0.7995
Logistic Regression with Bagged Trees	0.4597	0.8548	0.7686	0.8032	0.7284	0.7742

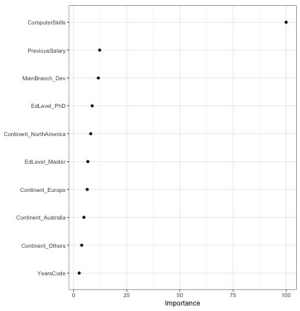
ComputerSkills
PreviousSalary
MainBranch_Dev
EdLevel_PhD
Continent_NorthAmerica
EdLevel_Master
Continent_Europe
Continent_Australia
Continent_Others
YearsCode
0
25
50
75
100
Importance

Figure 1

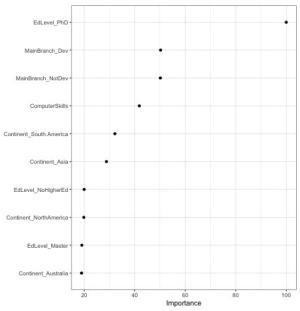
EdLevel_PhD
MainBranch_Dev
MainBranch_NotDev
ComputerSkills
Continent_South America
Continent_Asia
EdLevel_NoHigherEd
Continent_North America
EdLevel_Master
Continent_Australia
20
40
60
80
100
Importance

Figure 2