Job Attrition Report

Zervaan Borok

Table of Contents

Introduction

The purpose of this report is to analyze the performance of various support vector machine (SVM) models on job attrition data. The dataset used in the analysis that follows contains 1,470 observations of 31 variables. The objective of the models was to predict whether a given employee will suffer from job attrition based on their work and personal life characteristics described by the explanatory variables. Before diving into how the SVM classifier models were constructed, we will briefly review what SVM models actually are.

Support Vector Machines Explained

SVM algorithms are designed to identify a hyperplane situated in an N-dimensional space, where N corresponds to the number of features, that distinctly classifies the observations within a given dataset. For any given dataset, there are numerous hyperplanes that can be chosen in order to separate observations into two classes. However, the objective is to identify the hyperplane that contains the maximum margin, which is defined as the maximum distance between the observations of the two classes. This process enables future observations to be classified with greater confidence. Hyperplanes are used as decision boundaries that assist with classifying the observations. Observations which fall on either side of the hyperplane can be allocated to different groups. The observations located closest to the hyperplane are called support vectors; these observations help determine the position and orientation of the hyperplane and assist with the SVM construction. The values returned by an SVM model will fall within the range of [-1, 1]. If the output is equal to 1, the relevant observation is assigned to one class. If the output is equal to -1, the relevant observation is assigned to the other class. The hinge loss function is used to help maximize the margin (or distance) between observations and the hyperplane. In order to balance the margin maximization and the loss, a regularization parameter is added to the hinge loss function (Gandhi, 2018)[1]. The final SVM cost function is shown below (Liu, 2020)[2].

$$L(w) = \sum_{i=1} \underbrace{max(0, 1 - y_i[w^T x_i + b}_{\text{Loss function}}]) + \underbrace{\lambda||w||_2^2}_{\text{regularization}}$$

To find the gradients, partial derivatives are taken with respect to the existing weights. Subsequently, these gradients can be used to update the weights. When the model correctly predicts the class of any given observation, the gradient is only updated by the regularization parameter. Alternatively, when the model incorrectly predicts the class of any given observation, the gradient is updated by the entire loss function (Gandhi, 2018)[1].

Initial SVM Model

To construct the SVM classifier, the data set was split into training and testing sets. The training set was allocated 80% of the observations while the testing set was allocated 20%. Next, an SVM model was fitted to the training data set. The response variable was ‘Attrition’ and the predictor variables were all of the other variables available within the dataset. The kernel with which the SVM model was fitted with was the Radial Basis Function (RBF). The data was centered and scaled before being passed to the model. 10-fold cross validation was used in the model and the number of parameter values to try was set to 10. Following model execution, it was found that the highest accuracy was achieved when C was set to 2 and sigma was set to 0.00965926. The accuracy associated with these parameter values was 87.06% on the training dataset.

Linear SVM & KNN Models

For comparison, an SVM model with a linear kernel was fit to the training data set. Again, the data was centered and scaled before being passed to the model and 10-fold cross validation was used. Instead of asking the model to find the optimal value for the parameter C, the optimal C value as identified by the RBF SVM model was passed to the linear kernel. This model achieved a classification accuracy of 86.98% on the training dataset. Subsequently, a K-nearest-neighbor (KNN) model was built and passed the same centered and scaled training dataset. Once again, 10-fold cross validation was used and the number of parameter values to try was set to 10. The highest accuracy achieved by the KNN model on the training dataset was 84.68% when the parameter k was set equal to 7. We can see that the KNN model does not perform as well as either of the SVM models as evidenced by their accuracies.

Polynomial SVM Model

Finally, a fourth SVM model was constructed in which the kernel was set to ‘polynomial’. Again, this model was passed the same centered and scaled training dataset as the previous three models. 10-fold cross validation was used, but the number of parameter values to try was set to 4. The tuning parameters for this model were degree, scale, and C; their optimal values under the training dataset were found to be 1, 0.1, and 1 respectively. Following this, each model’s accuracy on unseen data was evaluated by passing the testing data set through each one in turn. The results of this analysis are shown in Table 1, below.

Table 1

SVM Poly Accuracy	SVM RBF Accuracy	SVM Linear Accuracy	KNN Accuracy
0.8881356	0.8813559	0.8677966	0.8440678

As evidenced by the table, the model that performs best on unseen data is the SVM model with the polynomial kernel. It should be noted that the difference in accuracy between the polynomial and RBF SVM models is only 0.67797%, but the polynomial model takes much longer to execute.

Employee Intervention

The first employee, Employee A, is a 48-year-old female who rarely travels for business and works in the sales department. Her hourly, daily, and monthly rates are £102, £1,202, and £19,479 respectively while her monthly income is £6,993; she has a college education in life sciences and lives 8 miles from the office. Her environment satisfaction is 'medium', her job involvement is 'high', and her job level is classified as 'level 2'. She is a 'sales executive' and her job satisfaction is rated as 'very high'. Her marital status is single, she has worked for 8 companies, is eligible for over-time pay, and received an 11% salary hike along with an 'excellent' performance rating. Her relationship satisfaction is 'low', she has no stock options, has worked for 8 years, had no 'training times' last year, and has a 'bad' work-life balance. She has been with the company for 6 years, 4 of which were spent in her current role, and 5 of which were spent under her current manager. Finally, she was just promoted this year. The second employee, Employee B, differs from Employee A. Employee B is a 38-year-old male with a Masters degree in life sciences who rarely travels for business, works in the sales department, and lives 9 miles from the office. His hourly, daily, and monthly rates are £82, £1,218, and £6,986 respectively while his monthly income is £3,407. He is a 'sales representative', which has a job level of 'level 1', with 'medium' job involvement and 'low' job satisfaction. He has worked at 7 companies, is not eligible for over-time pay, and received a 23% salary hike along with an 'outstanding' performance rating. His relationship satisfaction is 'medium', he has no stock options, has been working for 10 years, and had 4 training sessions last year. He has a 'better' work-life balance and has been with this company for 5 years, 3 of which were spent in his current role and with his current manager. He was also just promoted this year. Both employees are at risk for job attrition, but for different reasons. While Employee A has very high job satisfaction and an excellent performance rating, she has high job involvement, low relationship satisfaction, and a bad work-life balance. Sooner or later, low relationship satisfaction combined with a bad work-life balance will lead to problems in one's personal life. Once someone develops problems in their personal life due to poor work-life balance, their performance at work will start to deteriorate as their personal problems begin to leech into their professional lives (Ahmed, 2020)[3]. Employee A likely does not have time for much else asides work, so to make her job more attractive, it would probably be wise to either give her more paid time off or appoint another individual to share some of her responsibilities at work. Conversely, Employee B has medium job involvement and low job satisfaction, but medium relationship satisfaction and better work-life balance. It appears that Employee B has enough time for his personal life but is unsatisfied with his job. This could be due to a number of reasons, but it's probably because his job is not challenging enough for him. When people have to perform tasks below their skill set, they tend to become unsatisfied as the tasks are not fulfilling. Eventually, this will negatively impact performance at work as it is difficult to give 100% of your effort every day when you do not find your job very engaging (Quintini, 2011)[4]. Thus, to make his job more attractive, it would probably be wise to give him more responsibilities at work so that his job involvement progresses from 'medium' to 'high'.

References

1. Gandhi, R. (2018) *Support Vector Machine — Introduction to Machine Learning Algorithms.* Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Accessed 15 January 2022).

2. Liu, C. (2020) *A Top Machine Learning Algorithm Explained: Support Vector Machines (SVMs)*. Available at: https://www.vebuso.com/2020/02/a-top-machine-learning-algorithm-explained-support-vector-machines-svms/ (Accessed 15 January 2022).

3. Ahmed, A. (2020) *The Effects of Work & Life Imbalance*. Available at: https://work.chron.com/effects-work-life-imbalance-5967.html (Accessed 15 January 2022).

4. Quintini, G. (2011) *Over-Qualified or Under-Skilled: A Review of Existing Literature*. OECD Social, Employment and Migration Working Papers No. 121. doi: https://dx.doi.org/10.1787/5kg58j9d7b6d-en