Andrew Singh and Mike Johnson

OPAN 6603: Machine Learning II - Project 1

Executive Summary

Our team is providing an analytical summary to Canterra's leadership demonstrating the use of predictive models, K-Nearest Neighbors (kNN) and Support Vector Machine (SVM), to further understand their employee attrition. Our proposal demonstrates our methodology, summary of findings, and recommendations Canterra's leadership could consider implementing.

The key findings from our analysis include:

1) The k-Nearest Neighbors (kNN) model has moderate precision (0.75) but low recall (0.30), meaning that the model is highly precise, but misses many true positives.
2) The Support Vector Machines (SVM) model has both high precision (1.00) and high recall (0.98), meaning the model performs well making accurate predictions.
3) Both models lack actionable insights, unlike prior models developed using these data such as the logistic regression model.

Our recommendations for Canterra's leadership includes: 1) move forward with the SVM model for predicting attrition due to exceptional performance in making accurate predictions and 2) continue using the logistic regression model developed in the prior engagement due to its ease of interpretation.

Introduction

Our data analytics team supports Canterra's senior leadership by providing data-driven solutions to address challenges with managing talent, given an annual attrition rate of 15% (about 600 employees). One of Canterra's goals is to replace lost talent with new hires each year.

Leadership attributes this attrition to three main factors: 1) missed project timelines set by former employees, damaging relationships with consumers and partners, 2) the need to maintain adequately staffed departments with new talent, and 3) training incoming employees on their roles and company policies. Our goal is leveraging new predictive machine learning techniques to further understand causes of attrition.

Methodology and Data

We are using an internal Canterra employee dataset of 4,410 employees. The dataset provides employee-specific characteristics of each individual employee. The key variables we are interested in are attrition, job satisfaction, and total number of years worked by the employee as these characteristics were previously shown to impact Canterra's employee attrition.

For our methodology, we applied two supervised machine learning models, K-Nearest Neighbor (kNN) and Support Vector Machines (SVM).

K-Nearest Neighbor (kNN):
kNN predicts outcomes by comparing similarities, in this case similar attributes of employees to another employee. The algorithm identifies the “nearest neighbors" (known as NN) or observations we want to classify based on distance and assigns the most common attribute among them. The choice of how many “neighbors” (known as K) is a critical parameter affecting the model’s accuracy. The advantages of kNN are it is easy to use, works well for both predictions and classifications, and doesn’t require assumptions about the data. The shortcomings of kNN are that it can become computationally taxing as the dataset grows, slowing performance.

Support Vector Machines (SVM):
SVM classifies data by mapping observations, in this case, employees, into a multi-dimensional space, where each feature represents a dimension. The model identifies an optimal boundary known as a hyperplane or a linear line) that separates and classifies target classes, in this case whether employees were affected by attrition or not. The closest points to this boundary are called support vectors, which play a crucial role in defining the boundary or margin to make an implied decision. For complex patterns, SVM can adjust for flexibility, improving accuracy by applying a regularization parameter known as C. The benefits of using SVM is that it is generalizable, while shortcomings are to be considered on how to fine-tune and modify the kernel function with different parameters.

Key Findings
The models developed for this engagement demonstrated varying performance, with the SVM model performing significantly better at predicting attrition relative to the kNN model. The table below summarizes key metrics for the two models.

Model	Area Under the Curve (AUC)	Precision	Recall
k Nearest Neighbors	0.79	0.75	0.30
Support Vector Machines	0.97	1.00	0.98

k Nearest Neighbors Model
One of the key steps in building a kNN model is choosing the value of k. This value is key in classifying data points. To determine the optimal k value, cross validation was performed on k-values ranging from 2 to 40, and 2 was identified as the optimal value (Figure 2). At a k-value of 2, the model demonstrated moderate precision (0.75) and low recall (0.30). This means that the model is precise but misses many true positive cases.

Support Vector Machines Model

To determine the optimal boundary between categories; SVM models using linear, radial, and polynomial terms were trained. Out of the three trained, the radial model performed the best (Figure 2). The radial model has high precision (1.00) and high recall (0.98). This means that the model is highly accurate.

Recommendations

Our recommendation is dependent on Canterra leadership's intent with these models. For a purely predictive model, the SVM model is recommended. It's high precision and recall demonstrate its reliability in predicting attrition. However, if the goal is interpretability and understanding the contributing factors of attrition, we recommend that Canterra continue to reference the logistic regression model developed in our prior engagements. The logistic regression model is easy to interpret, translating complex findings into actionable insights for leadership.

Assuming the latter, we refer to what we recommended in the previous engagement to curb attrition:

- **Provide additional support for employees who need to travel.** This includes improved travel allowances and investing in wellness programs to help alleviate the stress associated with travel. In addition, technology solutions such as video conferencing should be considered to reduce the need for travel.
- **Develop targeted employee engagement strategies for single employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.
- **Conduct regular employee satisfaction surveys to identify pain points and areas for improvement**. Invest in programs that improve job satisfaction, such as employee recognition, career development opportunities, and initiatives to improve employee work-life balance.
- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.

Conclusion

The purpose is to build on the prior work to further understand root causes of Canterra's attrition. We found the SVM model is highly accurate, however, it lacks the business interpretability a stakeholder would need to understand for effective data-driven decision-making. We developed two models that are predictive in nature, but do not clearly identify what are attributing factors for Canterra's attrition. Our takeaway is that if the business objective is to clearly define the root cause of attrition, we recommend Canterra's leadership to reference the logistic regression model, due to its ease of interpretability. Our call to action is to refer to the logistic regression model and continue enhancing the model to understand influencing factors of Canterra's attrition.

Appendix

This condensed format ensures the report stays focused and digestible while maintaining

Figure 1: ROC Curves Comparison

The receiver-operating characteristic (ROC) plots for each model selected included below. The SVM model performed significantly better with an area under curve (AUC) of 0.97 vs the kNN model at 0.79.

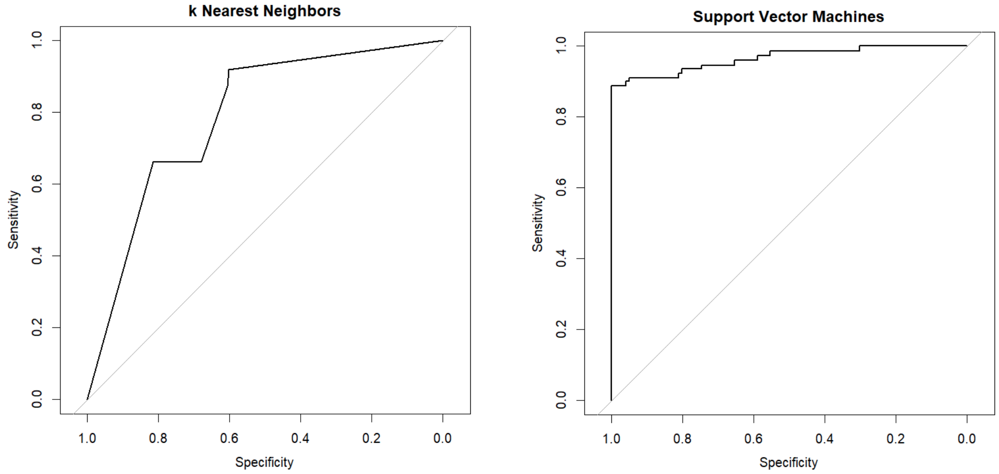


Figure 2: k-Cross Validation for the kNN Model

The visual below plots the receiver-operating characteristic (ROC) at various values of k. The optimal value for k is 2. As k increases, ROC decreases.

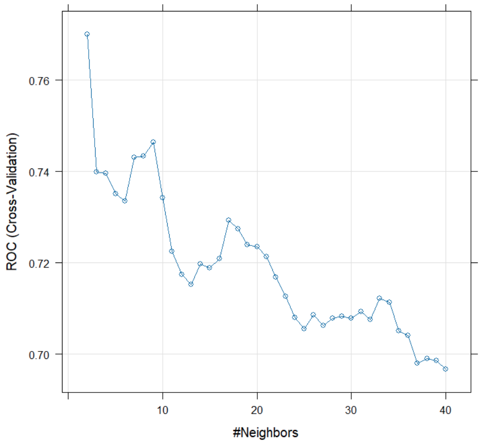
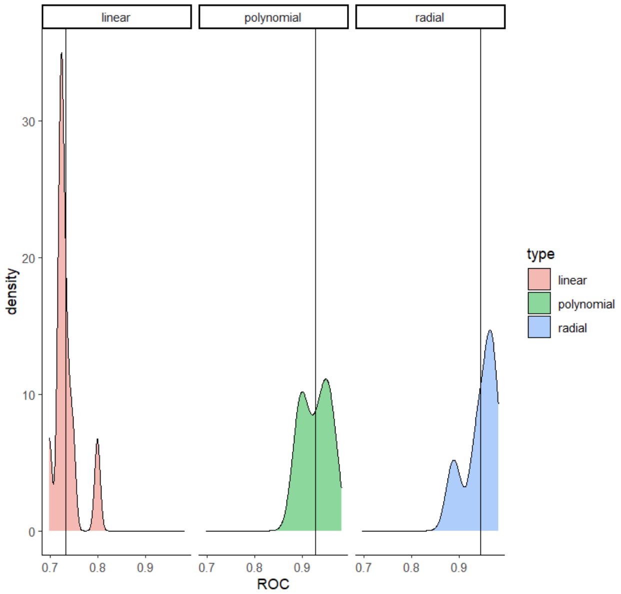


Figure 2: SVM Model Performance Comparison

The density plots below show the receiver-operating characteristic (ROC) during cross validation for each SVM Model. The radial model had the best performance during cross validation.


linear
polynomial
radial
density
ROC
type
linear
polynomial
radial