SDS 3786 Lab 7: La classification et la regression

Professor: Patrick Boily

Anthony Le, John Bute, Kelly Gao

Introduction

Our Canadian embassies host exceptional workers, but managing employee counts across numerous consulates is challenging, as we need to know the best number of employees needed to respond to the population's demands. To predict the average yearly employees per consulate, we applied classification and regression techniques. Average employee counts were categorized into quartiles: Very Low (0-25%), Low (25-50%), Medium (50-75%), and High (75-100%). For classification, models like Decision Trees, Naive Bayes, Neural Networks, and SVMs were used, while regression employed Decision Trees, Neural Networks, Linear Regression, and SVMs. Ensemble methods were applied to enhance accuracy. We reserved 20% of data for final evaluation, with the remaining split 80/20 for training and testing. To test the accuracy, we ran each model 20 times and took the average accuracy of all the runs. Here are our results:

Classification Results:

Decision Tree

Accuracy: 51.23%

Confusion Matrix:

Prediction/Actual	Very Low	Low	Moderate	High
Very Low	18	6	1	1
Low	13	32	31	13
Moderate	0	11	29	15
High	0	3	5	25

Naive Bayes

Accuracy: 44.33%

Confusion Matrix:

Prediction/Actual	Very Low	Low	Moderate	High
Very Low	25	30	23	9
Low	3	13	10	6
Moderate	1	7	25	12
High	2	2	8	27

Neural Network

Accuracy: 37.93%

Prediction/Actual	Very Low	Low	Moderate	High
Very Low	17	16	13	7
Low	12	28	28	15
Moderate	0	0	0	0
High	2	8	25	32

SVM

Accuracy: 51.23%

Prediction/Actual	VeryLow	Low	Moderate	High
Very Low	8	5	1	0
Low	23	57	29	15
Moderate	0	9	33	13
High	0	1	3	26

Ensemble of all classification models

Accuracy: 45.91%

Prediction/Actual	VeryLow	Low	Moderate	High
Very Low	6	2	3	1
Low	31	48	27	10
Moderate	3	10	18	11
High	0	6	35	46

Insights:

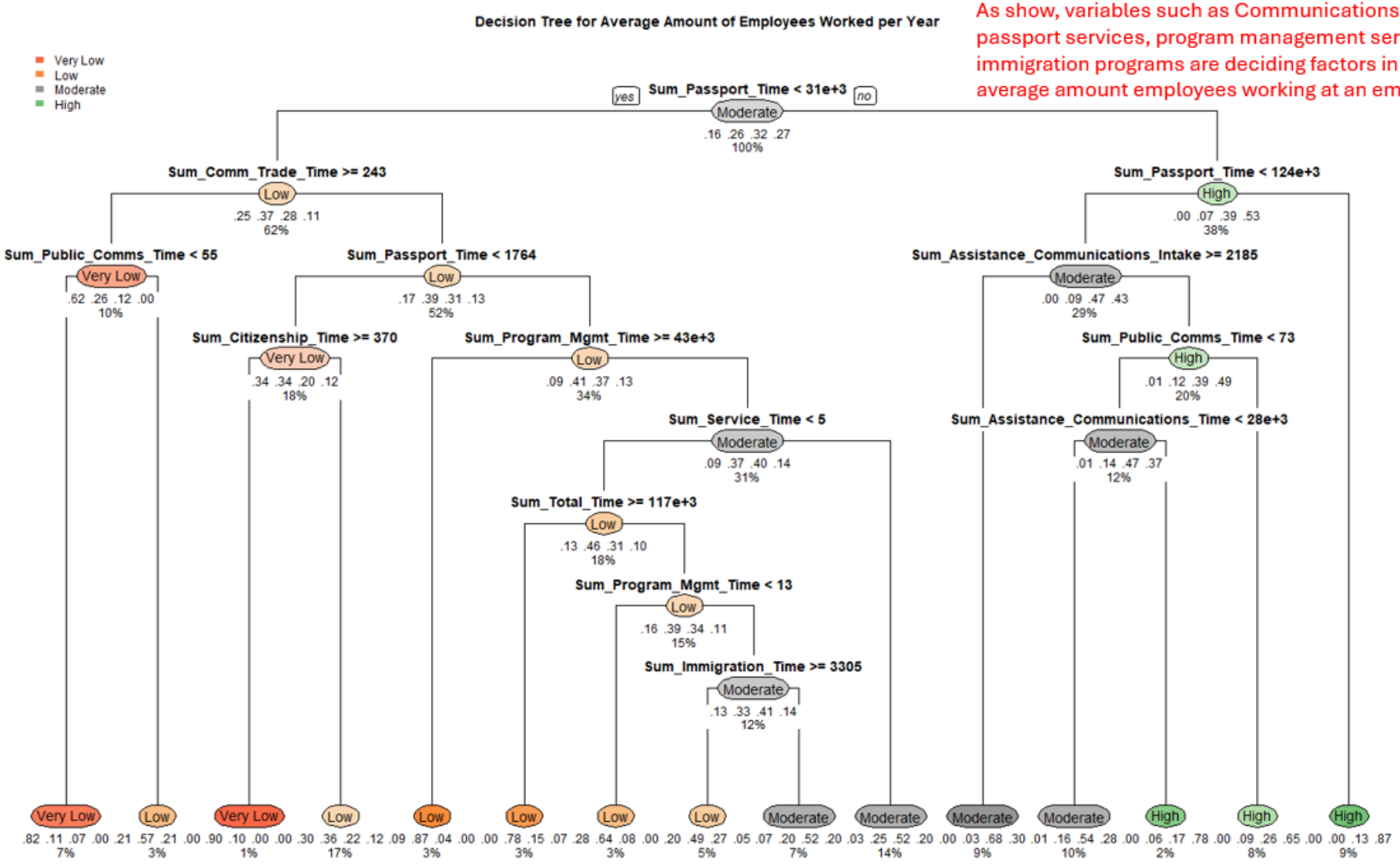
The decision Tree model performed reasonably well (51.23% accuracy). However, the misclassification of many observations as Moderate and Low suggests that the splits in the tree might not fully represent the complexity of our dataset. However, as we see in Plot 1, our decision tree tells us that our passport, citizenship, and communication services are important. This makes sense, as these categories represent some of our popular services (over 67.09% of the total amount of time spent working), so we must allocate more employees towards this.

However, Naive-Bayes struggled, as a 44.33% accuracy demonstrates a strong interdependence between variables. For example, intakes and TimeWorked on several categories were shown to have a positive correlation, thus violating the independence assumption of Naive Bayes.

Neural networks underperformed significantly due to insufficient tuning. We believe that a team dedicated to hyperparameter tuning and data augmentation could improve this performance, but due to our limited resources and the presence of better performing and more interpretable models, we won't explore further.

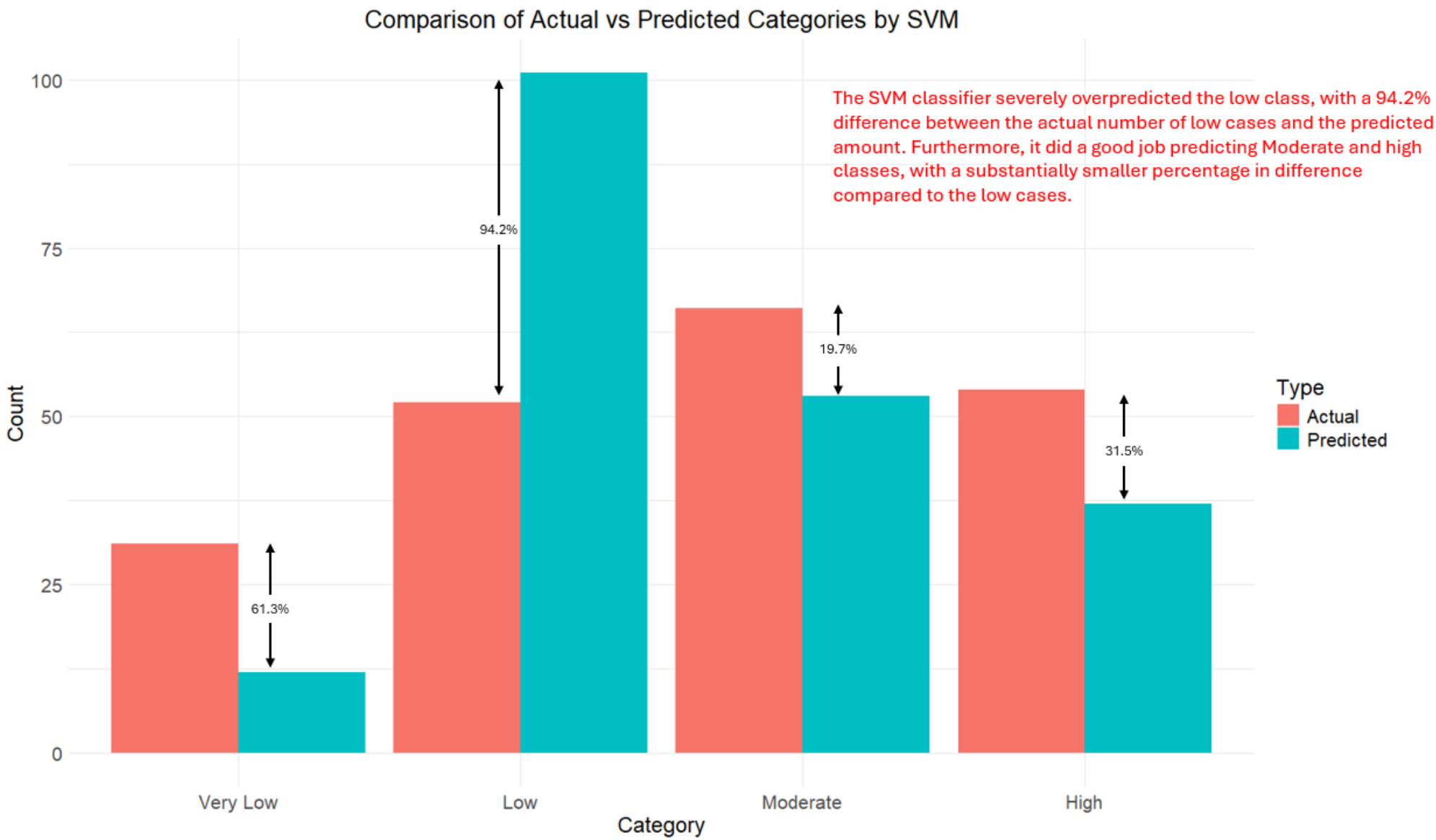
The SVM classifier performed exceptionally well, with an accuracy of 51.23%. However, as shown by plot 2, the SVM classifier overpredicted the low category and severely underpredicted the very low category. The former might be due to the low category having clearer decision boundaries than the rest, while the latter is probably due to dataset imbalance, as the number of cases of a very low average employee count is rarer than the other three

Finally, our ensemble model underperformed (45.91% accuracy). The cause of this is centered around the conflicted predictions of our individual models. A future avenue for this approach would be a weighting based on individual performance in order to reward right answers and punish wrong ones.



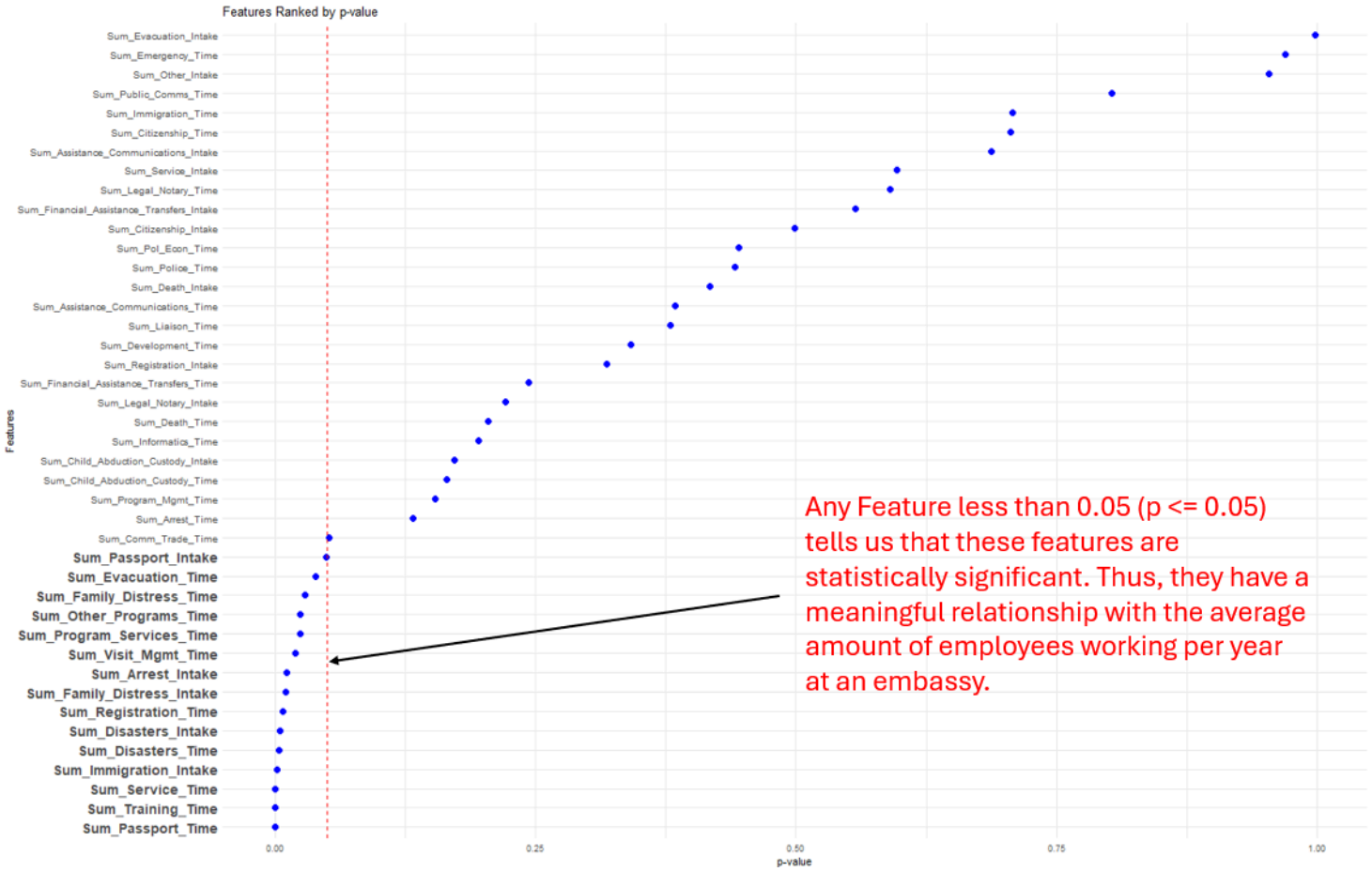
As show, variables such as Communications services, passport services, program management services, and immigration programs are deciding factors in classifying the average amount employees working at an embassy in a year

Plot 2: decision boundaries of scatter plot matrix


Comparison of Actual vs Predicted Categories by SVM
The SVM classifier severely overpredicted the low class, with a 94.2% difference between the actual number of low cases and the predicted amount. Furthermore, it did a good job predicting Moderate and high classes, with a substantially smaller percentage in difference compared to the low cases.
100
75
50
25
0
Count
94.2%
19.7%
31.5%
61.3%
Type
Actual
Predicted
Very Low
Low
Moderate
High
Category

Regression Insights

- **SVM Regression**: With the lowest RMSE (0.3088) and highest R^2 (0.4479), the model shows a good fit. However, it's only 2% better than linear regression and less interpretable. Thus, linear regression would be our preferred choice.
- **Linear Regression**: Decent performance (RMSE: 0.3147, R^2: 0.4240), but not as accurate as SVM. We plotted the p-values of our features to determine which ones are statistically important.
- **Decision Tree**: Moderate results (RMSE: 0.3265, R^2: 0.3820). May benefit from tuning or ensemble methods.
- **Neural Network**: Poor performance (RMSE: 0.4056, R^2: 0.0151), requires further hyperparameter tuning or different architecture.



Conclusion:

Our results demonstrate that passport, service, immigration, and communications programs are key predictors of average staffing needs at embassies as they appear as important features in our regression and decision tree model. We recommend prioritizing resources for embassies with high demand in these areas. If we want to predict if an embassy needs more employees or not, then we suggest using the decision tree as it excelled in classification or the SVM regression model to get the best possible estimate for the number of employees needed.