FYP Project Report

Data Sources

Data Processing

ML Models

Linear regression

Ridge regression

Cross-validation (CV)

* A model usually creates a prediction function from the same set of data, which might lead to a bias towards that dataset. This might produce inaccuracies when prompted with un-seen data (a situation called overfitting).
* Cross-validation solves this problem by using the K-fold strategy. The dataset is split into k smaller datasets and the model is trained and tested k times on k-1 folds and 1 fold respectively. Eg, a CV with 5 folds running on a dataset of 100 data will produce 20 data per fold.
* Cross-validation is computationally done using the cross\_val\_score helper function on the dataset by splitting the data, fitting a model, and computing the score for k number of consecutive times (with different splits each time). The Root Mean Squared Error (RMSE) is used as a performance metric.
* The results of each of the selected number of folds are documented below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K - folds | Linear Regression | | Ridge Regression | |
| Mean RSME | Standard Deviation | Mean RSME | Standard Deviation |
| 5 | 101.66 | 12.34 | 96.09 | 11.00 |
| 10 | 101.00 | 24.29 | 93.96 | 22.67 |
| 15 | 99.06 | 25.88 | 92.84 | 28.69 |
| 20 | 99.10 | 30.38 | 92.24 | 30.91 |

* Low Mean RMSE: This indicates that the model, on average, has a low prediction error, directly relates to the model's predictive power.
* Low Standard Deviation: This suggests that the model's performance is consistent across different subsets of the dataset, crucial for ensuring that the model is reliable and not just performing well on specific types of data.
* Considering the trade-offs between accuracy and consistency, a model with a balanced mean RMSE and standard deviation is chosen for the context of predicting PGTAs to hire. Hence, the optimal model is the ridge regression model with 5-fold cross-validation