Aditya Shah, C.J. D'Amico, Rahil Virani, Jesmyn Zhang
10/22/2024
Team Assignment 2
GB 656

Assumptions: The dataset contains number of workers with values of 0.5. We assumed that meant there could be part-time workers that were evaluated at 0.5 versus 1 for full-time workers. We are also assuming that any worker can be placed in program A and program B. This assumption means that we allow ourselves to use our decision rules with no bias to the workers. Lastly, we are assuming that every observation is independent of one another as correlation between observations would over-complicate our model and a simpler model is easier to implement.

Steps for Data Preparation and Analysis: The workers in progress variable was removed since it had many missing values. The date variable was removed because workday and quarter contained the information that was in the date variable. The idle men variable was also removed because idle time captures the same information. Idle time had several values of "0" so we dichotomized it for better analysis. We added a duplicate column for worker count so we could find profit per worker. Team, targeted productivity, and style changes were all variables we one-hot encoded to categorical variables to more easily find the correlation between values rather than have to find the correlation between continuous variables. We removed the duplicate column of worker count after splitting the data into a training test split of 80/20.

Client A

Model of Choice and Performance Assessment: We created four models: Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Support Vector Regressor (SVR). For each model, we found Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, Explained Variance, and Mean Absolute Percentage Error (MAPE). Each model was graphed to see which one had the smallest difference between Predicted and Actual Productivity to visually assess model accuracy. The lowest values for RMSE, MAE, and MAPE came from the Random Forest model, and the highest R^2 and Explained Variance were also in the Random Forest model. These reliable performance metrics as well as the closeness of the points on the graph to the 45-degree line show that the Random Forest model was the best model out of the four we tested.

Model Tuning: We arbitrarily decided to find the top 10 features that affected the model and only use those to fit the model as to prevent overfitting and reduce the complexity of the model. The top feature we found was incentive, which makes sense since incentives are meant to improve worker productivity and therefore has a direct correlation with the actual productivity variable. We also used a grid search with 5-fold cross-validation to tune our Random Forest model so we could train and strengthen the model on more random subsets of data. Combining these strategies improved the Random Forest R^2 by 4.4 percentage points to 0.489.

Decision Rules: Our decision rule for Client A is we only place workers whose predicted probability exceeds 0.58. This was based on the fact that a worker is only profitable if their productivity is greater than 0.583 since 70 (cost of worker) / 120 (productivity multiplier) = 0.583.

Model Profitability and Financial Value: The workers not placed had a profit of 0 and each of the workers placed had a profit listed under column model_profit. To sum up, this column gave us a model profit of $165,888.08. We found 80.19% of workers were placed in our model and running a random placement of 80.19% of workers gave a profit of $116,680.62. This gave us the financial value of the model as $49,207.46.

Client B:

Model of Choice and Performance Assessment: We created three models: Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). For each model, we printed classification reports to evaluate the accuracy of each model in correctly placing workers. Random Forest again had the highest accuracy with 86.25% as well as the highest precision and recall for placing and not placing workers indicating it balances the trade-off between false positives and false negatives better.

Model Tuning: We again arbitrarily decided to find the top 10 features that affected the model and only use those to fit the model as to prevent overfitting and reduce the complexity of the model. Using only the top 10 features for the model increased accuracy in the Random Forest model showing that the decreased model complexity helped its predictive capabilities and using all features overfit the model. The grid search also helped train the model and combining grid search and selecting features improved the Random Forest model accuracy to 0.90.

Decision Rules: For Client B, we calculated profits based on True Positives (TP) generating $180 per correct placement and False Positives (FP) incurring a loss of $70 per incorrect placement. True Negatives (TN) and False Negatives (FN) were not directly considered. We aimed to minimize FP and maximize TP by testing thresholds at intervals of 0.02. Our analysis revealed that the optimal threshold for maximum profit was 0.50. Therefore, our decision rule is to place workers only when their predicted probability of exceeding 0.80 is greater than 0.50.

Model Profitability and Financial Value: The optimal threshold for maximizing profit was found to be 0.50, yielding a maximum profit of $633,560. The average profit over five iterations of randomly assigning workers based on predicted productivity was $129,570, and when based on actual productivity, it was $171,420. The estimated model value for predicted productivity was $503,990, while for actual productivity, it was $462,140.

For Client B, the model achieved a maximum profit of **$633,560** and a financial value of **$462,140**, representing substantial gains. In comparison, Client A's model yielded a profit of **$165,888.08** and a financial value of **$49,207.46**, demonstrating a much lower profitability and financial impact