Based on the preliminary [data](https://www.kaggle.com/datasets/rhuebner/human-resources-data-set) comparison, factors such as Employee Absence, Employee Engagement (i.e. Employees’ emotional connection to the organization), and Employee satisfaction have displayed strong correlation with employee performance (rated using ordinal rank – Exceed, Fully Met, Need Improvement, PIP).

Race and Departments have shown some correlation with employee performance, but to a lesser degree.

**Ordinal logistic regression** was originally considered given the nature of the dependent variable – Employee Performance. Ordinal logistic regression is a type of logistic regression when the dependent variable is ordered (e.g., smallest to greatest) rather than it being binary.

[Ordinal Logistic Regression](https://www.youtube.com/watch?v=jWIJ7P1G9P4) requires the following [assumptions1](https://www.st-andrews.ac.uk/media/ceed/students/mathssupport/ordinal%20logistic%20regression.pdf):

* The dependent variable is measured on an ordinal level.
* One or more of the independent variables are either continuous, categorical or ordinal.
* No multi-collinearity - i.e. when two or more independent variables are highly correlated with each other.
* **Proportional Odds** - i.e. that each independent variable has an identical effect at each cumulative split of the ordinal dependent variable.
  + This assumption may be tested by applying logistic regression at varying *levels (e.g., Dissatisfied, Neutral, Satisfied)* of the dependent variable[**2**](https://peopleanalytics-regression-book.org/ord-reg.html)

However, little resources were available on Excel to accurately test the assumptions of ordinal logistics regression and apply the method. Therefore, [**standard logistic regression**](https://www.ibm.com/think/topics/logistic-regression) was used in lieu of ordinal logistic regression.

Employee performance was simplified to either **(0 = Optimal Performance)** or **(1 = Suboptimal Performance)**. Performance expectations exceeding/fully met were categorized as optimal, and Employees needing improvement/on PIP were categorized as suboptimal.

Logistic regression model was created using Excel Solver and confusion matrix was used to obtain the optimal cut-off point as well as the ROC curve[**3**](https://www.youtube.com/watch?v=r9bIRMTZ6eM)**,**[**4**](https://www.youtube.com/watch?v=_rEW5ce78es). However, quick glance at the y-intercept value raised an eyebrow as the log odds indicated a value of 10.35. Converting the log odds to a probability [(exp(b0)/(exp(b0)+1))](https://www.bookdown.org/rwnahhas/RMPH/blr-interp.html), you would get that the probability of an employee having a suboptimal performance score at reference level (i.e., all other predictors unaccounted for) was 0.9997%. Initially, I thought the assessment of the y-intercept to be wildly unexpected and tried searching for faults within the data. That’s when I discovered **Penalized likelihood** or **Firth method**[**5**](https://statisticalhorizons.com/logistic-regression-for-rare-events/)**,**[**6**](https://academicweb.nd.edu/~rwilliam/stats3/RareEvents.pdf) typically used in cases where the positive outcome (or outcome of interest) in a logistic regression is extremely **rare** (2% rare). Luckily, my data was considered not extremely rare with approximately (9.96%) observations in the dataset representing positive outcome – i.e., suboptimal performance.

Model diagnostic was completed on the model to ensure model assumptions were satisfied. Logistic regression model was constructed using employee performance score against employee satisfaction, engagement survey results and count of absence.

Linearity assumption was completed by plotting the comparison between the continuous predictors against their logit outcomes. The following plot illustrates relatively linear relationship for **employee engagement survey** **results** and **satisfaction level**.

A graph of a graph of a logit

AI-generated content may be incorrect.

It’s important to note that employee satisfaction was treated as a continuous variable despite having ordinal characteristics (i.e., 1 = very dissatisfied, 2 = dissatisfied, 3 = neutral, 4 = satisfied, 5 = very satisfied). This was a deliberate decision made to maintain ascending order level.

It is important to note that since positive outcome (y=1) is assigned to employee’s odds of scoring poorly on the performance assessment, any increase in logit outcome implies higher odds of poor performance.

The plot results then make intuitive sense as one would expect a performance level to decrease as employee satisfaction and engagement levels drop.

The plot of absence is much less linear thus suggesting weaker association. This outcome matches the statistical insignificance absence has on the overall performance score (p-value of 0.8716). Absence was initially considered to be a significant predictor of employee performance level when the mode of absence between performance levels revealed significantly high occurrence of absence among poorly performing employees. This decision was supported by the fact that the distribution of employee absence at varying levels exhibited a multi-modal characteristic and therefore average would be an inappropriate measure of central tendency.

**A group of blue bars

AI-generated content may be incorrect.**

Upon closer inspection, however, it’s possible that absence levels are recorded lower among highly performing employees simply because of the sheer volume of data relative to their poorly performing counterparts.

Further approach may aim to determine its significance in the model.

Next, the model was tested for possible influential outliers. Cook’s distance was plotted to highlight the top 3 outliers.

A graph with numbers and lines

AI-generated content may be incorrect.

However, not all outliers are deemed influential, and thus a standard residual plot was used to highlight the influential outlier. R-code results deemed the top (highlighted) value as an influential outlier. The problematic outlier was then removed from the data to satisfy the assumption.

A graph with red and blue dots

AI-generated content may be incorrect.A graph with dots and numbers

AI-generated content may be incorrect.

Lastly, multicollinearity was tested to ensure no two predictors were too similarly correlated. This test was conducted by determining the VIF of all predictors and all predictors were not considered problematic[**7**](https://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/).

Considering the model mostly satisfies the assumptions, coefficients of the logistic regression were reviewed. As previously mentioned, at reference (base probability of low performance score assuming 0 employee satisfaction, 0 on engagement survey, 0 absence) is 99.96%. This makes sense as an employee with zero employee satisfaction and organizational loyalty are likely less incentivized to perform well.

Each unit increase in employee satisfaction was 0.47 times more likely to perform poorly. This means each increase in employee satisfaction will increase performance rating.

Each unit increase in employee engagement was 0.04 times more likely to perform poorly. Since each increase in employee engagement significantly reduces the odds of performing poorly, employee engagement has the strongest effect to performance out of all 3 predictors.

Each unit increase in absence was 0.99 times more likely to perform poorly. Since the coefficient is so close to 1, it means that each unit change in absence will have negligible change on performance.

Both employee satisfaction and engagement were considered **statistically significant** (p-values < 0.05) while absence was considered statistically insignificant.

Another model was fitted including secondary predictors (race, department), but race and departments were not considered statistically significant predictors of performance score (p-values > 0.05).

With the model diagnostic completed and coefficients identified, let’s test the accuracy of the model.

Prediction matrix was used to compare the number of true positives/negatives to false positives/negatives.

At 0.5 cutoff (as default), **99.29**% of true negative and **73.33**% of true positives were identified. Furthermore, the model had **96.8**% overall accuracy (TP + TN / Total), 91.7% precision (percentage of correctly identified TP by the model), and **73.3**% sensitivity/recall (percentage of correctly identified positive outcome by the model compared to the actual count of positive outcomes – i.e. TP).

Prediction matrices at different threshold levels were used to identify the optimal threshold.

Threshold of 0.4 was identified as the best threshold with **96.5%** **accuracy**, **82.8% precision**, and **80.0% sensitivity**. Based on the plot above, accuracy reaches its maximum at threshold 0.2 where it is maintained across rest of the cutoff threshold. Precision is also seen to increase as threshold increases but plateaus at around 0.6. Recall or sensitivity starts as 100% (meaning all datapoints are considered positives) and maintains a slight decrease between 0.1 to 0.4 after which the TP rate decreases significantly. 0.4 was identified as the optimal threshold since it maintains sufficiently high accuracy and precision without losing its TP rate significantly. Threshold after 0.4 sees significant decrease in the TP/sensitivity meaning more positive outcomes may be inaccurately categorized as negatives (false negative).

ROC curve provides a visual summary of all confusion matrices created at each cutoff/threshold. By comparing the **true positive rate** (proportion of employees correctly categorized as poorly performing) with the **false positive rate** (proportion of employees falsely categorized as poorly performing), ROC curve shows which threshold provides the most appropriate balance between TP/FP rate.

With regards to the model date, the ROC curve is shown as follows.

The plot shows how TP rate plateaus around the 80% mark and each increase of TP rate from this point onward leads to significant increase of FP rate. Threshold of 0.4 provides the most accurate **proportions of correctly classified employees** **(80%)** with **least amount of incorrectly classified employee** **(1.8%).**

A quick test was conducted to explore the effect of removing the influential outlier. Below table shows the outcome of confusion matrices at various threshold levels **prior to removing the influential outlier**.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cutoff | Accuracy | Precision | Recall |  | Cutoff | FP | TP |
|  | 96.1% | 91.3% | 67.7% |  |  | 0.7% | 67.7% |
| 0 | 10.0% | 10.0% | 100.0% |  | 0 | 100.0% | 100.0% |
| 0.1 | 86.2% | 40.3% | 80.6% |  | 0.1 | 13.2% | 80.6% |
| 0.2 | 92.3% | 58.5% | 77.4% |  | 0.2 | 6.1% | 77.4% |
| 0.3 | 94.5% | 70.6% | 77.4% |  | 0.3 | 3.6% | 77.4% |
| 0.4 | 96.5% | 85.7% | 77.4% |  | 0.4 | 1.4% | 77.4% |
| 0.5 | 96.1% | 91.3% | 67.7% |  | 0.5 | 0.7% | 67.7% |
| 0.6 | 96.1% | 100.0% | 61.3% |  | 0.6 | 0.0% | 61.3% |
| 0.7 | 95.2% | 100.0% | 51.6% |  | 0.7 | 0.0% | 51.6% |
| 0.8 | 94.2% | 100.0% | 41.9% |  | 0.8 | 0.0% | 41.9% |
| 0.9 | 93.2% | 100.0% | 32.3% |  | 1 | 0.0% | 0.0% |

The second table below shows the outcome **after removing the outlier**.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cutoff | Accuracy | Precision | Recall |  | Cutoff | FP | TP |
|  | 96.8% | 91.7% | 73.3% |  |  | 0.7% | 73.3% |
| 0 | 9.7% | 9.7% | 100.0% |  | 0 | 100.0% | 100.0% |
| 0.1 | 88.4% | 44.4% | 80.0% |  | 0.1 | 10.7% | 80.0% |
| 0.2 | 93.9% | 64.9% | 80.0% |  | 0.2 | 4.6% | 80.0% |
| 0.3 | 94.8% | 70.6% | 80.0% |  | 0.3 | 3.6% | 80.0% |
| 0.4 | 96.5% | 82.8% | 80.0% |  | 0.4 | 1.8% | 80.0% |
| 0.5 | 96.8% | 91.7% | 73.3% |  | 0.5 | 0.7% | 73.3% |
| 0.6 | 97.1% | 100.0% | 70.0% |  | 0.6 | 0.0% | 70.0% |
| 0.7 | 95.8% | 100.0% | 56.7% |  | 0.7 | 0.0% | 56.7% |
| 0.8 | 95.2% | 100.0% | 50.0% |  | 0.8 | 0.0% | 50.0% |
| 0.9 | 93.9% | 100.0% | 36.7% |  | 1 | 0.0% | 0.0% |

The model without the influential outlier is seen to better encapsulate the number of TP than its former, thereby confirming that addressing the outlier produces more accurate results.