# Project Analysis - Absenteeism

**Sagar Bansal**

## SECTION I: ANALYSIS SUMMARY

We discovered that Salary and whether a raise was given or not to the employee are crucial in identifying employees who are most likely to use a leave day. To a great surprise, employment status and age does not significantly affect the likelihood of the response. It is in contrast to the general belief that part-time employees are more likely to abuse leave policy as they are less loyal/dependent on the company.

A good estimation of accuracy for prediction based on the data will be 62.4%. Although we can modify the model to reach up to 80% accuracy, it would drastically reduce the percentage of accurately identified employees who actually will use the leave policy. Yes, we can tune the model so that at most 10% of the people who don't use the leave policy are incorrectly predicted by the model to use the leave policy. In that case, the people that will be identified to use policy who actually will use the leave policy in three months will range from 0.0% to 17.5%.

Besides this, client should keep in mind that response may be affected by the weather and season of that quarter. The tendency to take leave may have a seasonal pattern. Also, data was highly imbalanced which could give us a false sense of accuracy. Finally, additional features such as Married or not and Reason for leave may have added more to the analysis.

## SECTION II: APPENDIX

When we summarized the given data, we found that the average value of Salary and Age variables in the given data are roughly 47519.64 USD and 35, respectively. We made a bar graph on our response variable i.e., tookLeave and found that 20% of the employees took a leave which is significantly high considering the short period length i.e., three months. In addition, the box plot of age categorized by tookLeave didn’t show any considerable difference between the two categories. Although the minimum and maximum age is different for both categories, the mean and median are roughly same. Hence, we decided to not include age as a predictor in our model.

Since our response variable is a qualitative variable and we had to classify whether each employee is likely to take a leave or not, we used Logistic Regression in this case. The model had tookLeave as dependent variable and salary, employmentStatus & raiseOrPromo as independent variables. We found that the model is useful as the ROC curve is always above the diagonal line (random guessing) and AUC value is 0.64 which is greater than 0.5. For the model validity, we used Hosmer and Lemeshow Goodness-of-Fit Test to test the null hypothesis that the logistic model built accurately describes the data at 5% significance level and found that the p-value is 0.7827 which is greater than 0.05. Hence, we fail to reject the null hypothesis i.e., there is no evidence that the model does not fit well. Moreover, we didn’t notice any pattern in Pearson Chi-Square Residuals vs Case Number plot. Thus, we can say that there is no dependence issue in this model.

Using Wald Chi-squared test, we revealed that Salary (0.0001 < 0.05) and RaiseorPromo (0.0047 < 0.05) are statistically significant whereas employment status (0.2515 > 0.05) does not significantly affect the likelihood of the response at 5% significance level. incorrectly predicted by the model to use the leave policy. For the model accuracy, we chose pprob cut-off of 0.22 which gives us both Specificity (64.2) and Sensitivity (55.3) greater than 50%.

**Figure 1: Bar Graph - Fsrequency vs tookLeaave**

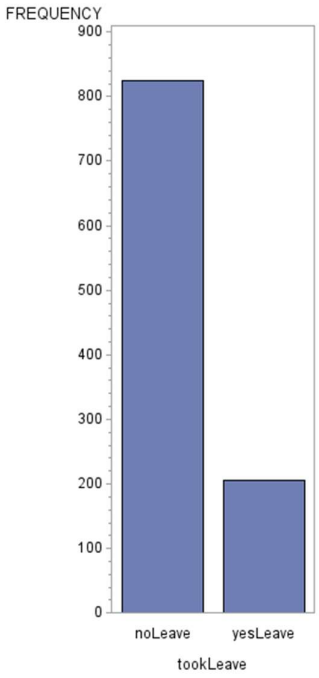



**Figure 2: ROC Curve – Sensitivity vs Specificity**

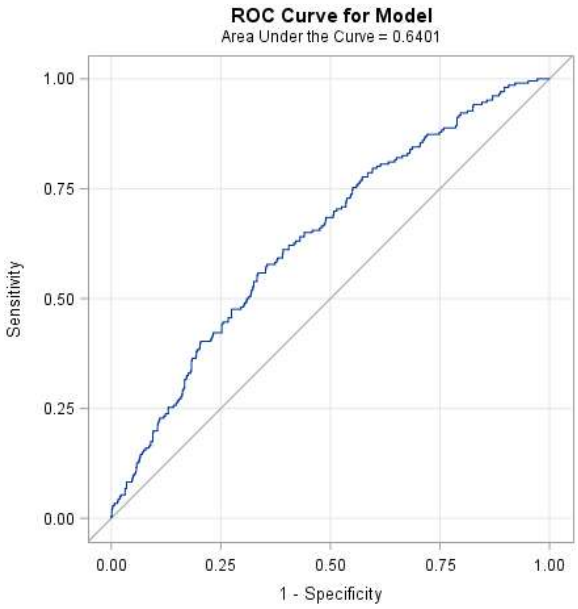



**Figure 3: Box plot – tookLeave vs Age**

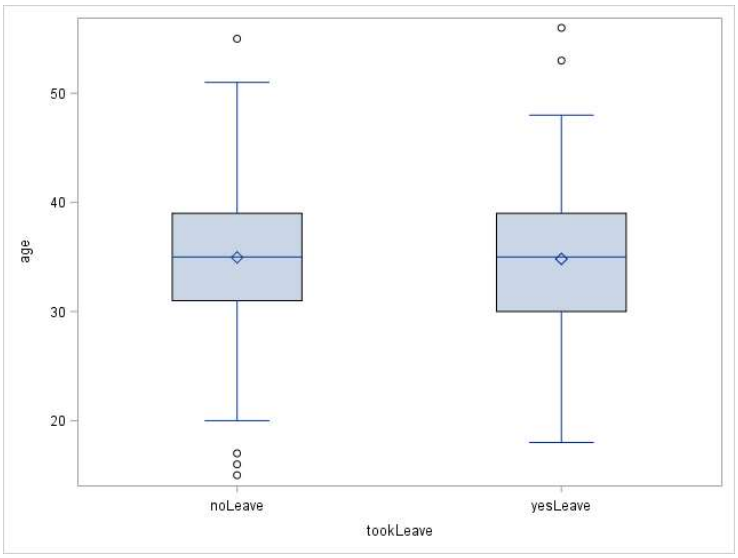
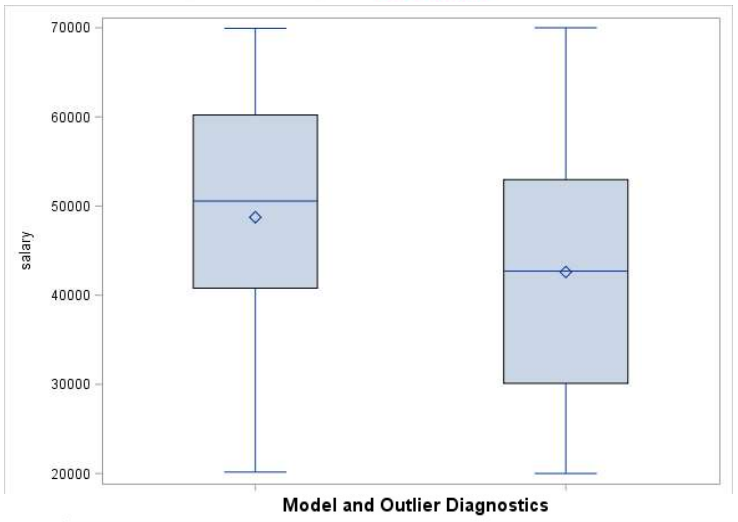



**Figure 4: Box plot – tookLeave vs Salary**




**Model and Outlier Diagnostics**

**Figure 5: Pearson Chi-Square Residual vs Case Number**

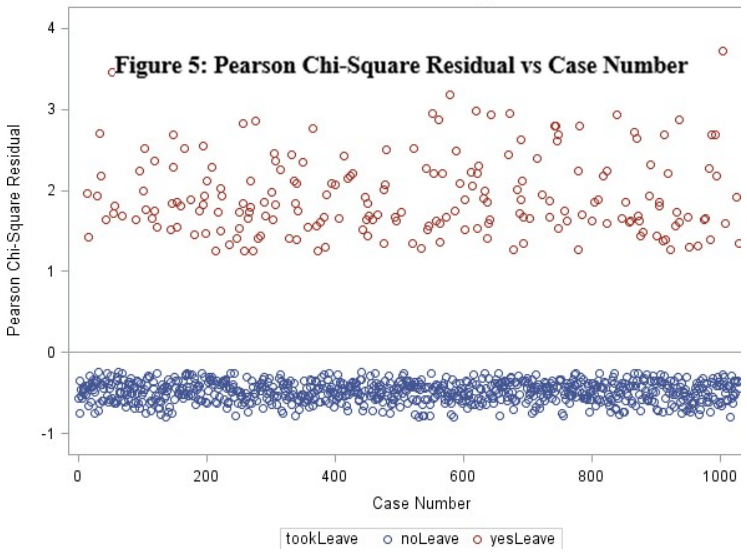


**Table 1 & 2: Numerical Summary of Salary**

The UNIVARIATE Procedure
Variable: age

| Moments | | | |
|---|---|---|---|
| N | 1030 | Sum Weights | 1030 |
| Mean | 34.9330097 | Sum Observations | 35981 |
| Std Deviation | 6.13631812 | Variance | 37.6544001 |
| Skewness | -0.0629494 | Kurtosis | 0.0754738 |
| Uncorrected SS | 1295671 | Corrected SS | 38746.3777 |
| Coeff Variation | 17.5659589 | Std Error Mean | 0.1912006 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 34.93301 | Std Deviation | 6.13632 |
| Median | 35.00000 | Variance | 37.65440 |
| Mode | 36.00000 | Range | 41.00000 |
| | | Interquartile Range | 8.00000 |

**Table 3 & 4: Numerical Summary of Age**

The UNIVARIATE Procedure
Variable: salary

| Moments | | | |
|---|---|---|---|
| N | 1030 | Sum Weights | 1030 |
| Mean | 47519.6408 | Sum Observations | 48945230 |
| Std Deviation | 14214.7566 | Variance | 202059304 |
| Skewness | -0.3046421 | Kurtosis | -1.0287839 |
| Uncorrected SS | 2.53378E12 | Corrected SS | 2.07919E11 |
| Coeff Variation | 29.9134344 | Std Error Mean | 442.915434 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 47519.64 | Std Deviation | 14215 |
| Median | 48917.00 | Variance | 202059304 |
| Mode | 28943.50 | Range | 49993 |
| | | Interquartile Range | 24712 |

**Table 5 & 6: Analysis of Effects and Analysis of Maximum Likelihood Estimation**

| Type 3 Analysis of Effects | | | |
|---|---|---|---|
| Effect | DF | Wald Chi-Square | Pr > ChiSq |
| salary | 1 | 14.7173 | 0.0001 |
| employmentStatus | 1 | 1.3150 | 0.2515 |
| raiseOrPromo | 1 | 7.9778 | 0.0047 |

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | | 1 | -0.2379 | 0.3746 | 0.4032 | 0.5254 |
| salary | | 1 | -0.00004 | 0.000011 | 14.7173 | 0.0001 |
| employmentStatus | FullTime | 1 | 0.3696 | 0.3223 | 1.3150 | 0.2515 |
| employmentStatus | PartTime | 0 | 0 | . | . | . |
| raiseOrPromo | no | 1 | 0.6245 | 0.2211 | 7.9778 | 0.0047 |
| raiseOrPromo | yes | 0 | 0 | . | . | . |

**Table 7: Classification Table**

| Classification Table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct | | Incorrect | | Percentages | | | | |
| Prob Level | Event | Non-Event | Event | Non-Event | Correct | Sensi-tivity | Speci-ficity | False POS | False NEG |
| 0.040 | 206 | 0 | 824 | 0 | 20.0 | 100.0 | 0.0 | 80.0 | . |
| 0.060 | 206 | 7 | 817 | 0 | 20.7 | 100.0 | 0.8 | 79.9 | 0.0 |
| 0.080 | 204 | 44 | 780 | 2 | 24.1 | 99.0 | 5.3 | 79.3 | 4.3 |
| 0.100 | 200 | 75 | 749 | 6 | 26.7 | 97.1 | 9.1 | 78.9 | 7.4 |
| 0.120 | 183 | 170 | 654 | 23 | 34.3 | 88.8 | 20.6 | 78.1 | 11.9 |
| 0.140 | 172 | 245 | 579 | 34 | 40.5 | 83.5 | 29.7 | 77.1 | 12.2 |
| 0.160 | 158 | 325 | 499 | 48 | 46.9 | 76.7 | 39.4 | 76.0 | 12.9 |
| 0.180 | 145 | 395 | 429 | 61 | 52.4 | 70.4 | 47.9 | 74.7 | 13.4 |
| 0.200 | 131 | 465 | 359 | 75 | 57.9 | 63.6 | 56.4 | 73.3 | 13.9 |
| 0.220 | 114 | 529 | 295 | 92 | 62.4 | 55.3 | 64.2 | 72.1 | 14.8 |
| 0.240 | 96 | 586 | 238 | 110 | 66.2 | 46.6 | 71.1 | 71.3 | 15.8 |
| 0.260 | 75 | 655 | 169 | 131 | 70.9 | 36.4 | 79.5 | 69.3 | 16.7 |
| 0.280 | 47 | 707 | 117 | 159 | 73.2 | 22.8 | 85.8 | 71.3 | 18.4 |
| 0.300 | 36 | 745 | 79 | 170 | 75.8 | 17.5 | 90.4 | 68.7 | 18.6 |
| 0.320 | 30 | 763 | 61 | 176 | 77.0 | 14.6 | 92.6 | 67.0 | 18.7 |
| 0.340 | 18 | 781 | 43 | 188 | 77.6 | 8.7 | 94.8 | 70.5 | 19.4 |
| 0.360 | 11 | 798 | 26 | 195 | 78.5 | 5.3 | 96.8 | 70.3 | 19.6 |
| 0.380 | 6 | 815 | 9 | 200 | 79.7 | 2.9 | 98.9 | 60.0 | 19.7 |
| 0.400 | 0 | 824 | 0 | 206 | 80.0 | 0.0 | 100.0 | . | 20.0 |

**Table 8: Hosmer and Lemeshow Goodness-of-Fit Test**

| Hosmer and Lemeshow Goodness-of-Fit Test | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 4.7623 | 8 | 0.7827 |