**Employee Attrition: The Great Resignation**

Team 2 Project

Professor Lin Hao

BYGB 7967-001: Data Mining for Business

December 6, 2022

**Introduction**

The most recent US economic history has progressed and grown since the Great Recession in 2009. The US transitioned from one presidency to another, the economic cycle continues to make consistent gains. As we entered the fall of 2019, China saw the emergence of COVID-19 which slowly spread through the Asian continent and then swiftly spread throughout the world. The spread of the virus triggered massive lockdowns and restrictions, during which most people began to work from home. The lockdowns and restrictions forced adjustment for the US workforce to work from home, which in turn created the unintended consequence of employees preferring to work from home. The lockdowns also gave people time to reflect on their priorities, which in turn caused people to look at their careers and evaluate whether those careers allowed them the freedom to pursue outside interests. Thus, many chose to terminate their business relationship with their respective employers. The high attrition caused many Human Resource (HR) departments to study the causes of the increased attrition rate and possible areas of improvement. For our project, we (the HR Compensation Analyst Team) have decided to approach this as a scenario situation in which we have been tasked by our Chief People Officer to study attrition within our company, a pharmaceutical company. Former and current employees were given a survey as to how they responded to certain duties in their current position and company policies.

The content of this report was produced using the DIDA framework, which consists of Data, Insights, Decisions, and Advantages. Our goal is to determine the 5 highest ranking issues which may influence an employee’s decision to seek other employment opportunities.

**Dataset Description**

The dataset contains data on 1471 fictional employees and 33 independent variables created by IBM data scientists. It contains 35 predictor variables in total, which consists of 17 numeric variables and 18 categorical predictors. The entire dataset is ex-ante and has “Attrition” as the dependent variable for our project. The variables have been organized into the table below:

| **Numerical Variables** | | **Categorical Variables** | |
| --- | --- | --- | --- |
| Age | PerecentSalaryHike | BusinessTravel | JobInvolvement |
| DailyRate | TotalWorkingYears | Gender | JobLevel |
| DistanceFromHome | TrainingTimesLastYear | MaritalStatus | JobSatisfaction |
| HourlyRate | YearsAtCompany | OverTime | PerformanceRating |
| MonthlyIncome | YearsInCurrentRole | StockOptionLevel | RelationshipSatisfaction |
| MonthlyRate | YearsSinceLastPromotion | Department | WorkLifeBalance |
| NumCompanies-  Worked | YearswithCurrManager | Education | Over18 |
| EmployeeCount | EmployeeNumber | EducationField | JobRole |
| StandardHours |  | Environment-  Satisfaction |  |

Several predictors were found to be redundant upon review the data and were removed. Those variables are EmployeeCount, EmployeeNumber, Over18, and StandardHours. Finally, the dependent variable is Attrition which we will predict as binary. The techniques we used in data mining the surveys included the Logistic Regression Model and Classification Tree.

**Insights and Evaluation Models**

1. **Logistic Regression Model**

Since our task is to predict the probability (which should range from 0 to 1) of a person leaving the company, we used logistic regression for this project as the technique for this project. With the prepared predictors and the dependent variable in the dataset, we will come out with multiple models and test for the best performance and parsimonious structure with cross-validation.

The model gave us the following coefficients.

| Age | -0.337719 | JobInvolvement\_1 | 1.690443 |
| --- | --- | --- | --- |
| DailyRate | -0.201618 | JobInvolvement\_2 | 0.080123 |
| DistanceFromHome | 0.501395 | JobInvolvement\_4 | -0.390306 |
| HourlyRate | 0.159242 | JobLevel\_2 | -1.322447 |
| MonthlyIncome | -0.572078 | JobLevel\_3 | 0.028852 |
| MonthlyRate | 0.013583 | JobLevel\_4 | -0.275973 |
| NumCompaniesWorked | 0.401974 | JobLevel\_5 | 1.709461 |
| PercentSalaryHike | -0.110726 | JobRole\_Healthcare | -0.624085 |
| TotalWorkingYears | -0.350199 | JobRole\_Human | 0 |
| TrainingTimesLastYear | -0.198439 | JobRole\_Laboratory | 0.124704 |
| YearsAtCompany | 0.573594 | JobRole\_Manager | -1.150994 |
| YearsInCurrentRole | -0.481707 | JobRole\_Manufacturing | -0.455882 |
| YearsSinceLastPromotion | 0.599446 | JobRole\_Research | -1.452757 |
| YearsWithCurrManager | -0.620476 | JobRole\_Research | -1.034361 |
| BusinessTravel\_Non-Travel | -1.404769 | JobRole\_Sales | 0.029308 |
| BusinessTravel\_Travel\_Frequently | 0.894681 | JobSatisfaction\_1 | 0.900026 |
| Department\_Human | -0.626729 | JobSatisfaction\_2 | 0.17832 |
| Department\_Sales | 0.770982 | JobSatisfaction\_3 | 0.502171 |
| Education\_1 | -0.402891 | MaritalStatus\_Divorced | -0.050912 |
| Education\_2 | -0.178786 | MaritalStatus\_Single | 0.291418 |
| Education\_4 | 0.030556 | OverTime\_Yes | 1.941412 |
| Education\_5 | 0.110795 | PerformanceRating\_4 | 0 |
| EducationField\_Human | 1.295719 | RelationshipSatisfaction\_1 | 0.81512 |
| EducationField\_Marketing | 0.5014 | RelationshipSatisfaction\_2 | 0.062691 |
| EducationField\_Medical | -0.060461 | RelationshipSatisfaction\_4 | 0.149312 |
| EducationField\_Other | 0.246403 | StockOptionLevel\_1 | -1.026333 |
| EducationField\_Technical | 1.039329 | StockOptionLevel\_2 | -1.378221 |
| EnvironmentSatisfaction\_1 | 1.355859 | StockOptionLevel\_3 | -0.383311 |
| EnvironmentSatisfaction\_2 | 0.107769 | WorkLifeBalance\_1 | 1.622664 |
| EnvironmentSatisfaction\_4 | 0.085949 | WorkLifeBalance\_2 | 0.519426 |
| Gender\_Male | 0.333894 | WorkLifeBalance\_4 | 0 |

The top five predictors from our model came out to be OverTIme\_Yes, JobInvolvement\_2, WorkLifeBalance, JobLevel\_2, and BusinessTravel\_Non-Travel. While contrarily our bottom 5 predictors were JobLevel\_5, RelationshipSatisfaction\_1, MonthlyIncome, YearsSinceLastPromotion, and YearsWithCurrManager.

The unexpected values derived from the DailyRate, HourlyRate, and MonthlyRate do not have a high correlation to Attrition occurring. Two job roles are highly correlated with whether a person leaves or not. They are a laboratory technician and research scientist. While another finding was that stock options showed to decrease the probability of a person separating from the company. This model that we built gave us an accuracy rate of 84 percent.

One problem when using the logistic regression model is the difficulty for the company to determine what set of conditions would impact the employees’ decisions. Knowing the top factors that cause people to leave is good, but the HR department would spend much more time and energy if they only looked for one factor at a time. Luckily, the Classification Tree Model can demonstrate qualitative and visualized insights and display a combination of conditions for HR’s reference.

1. **Classification Tree Model**

The Decision Tree Model provided a more straightforward graph for us to observe. The model has a hierarchical structure, which makes the decision-makers easy to follow the logic of the outcome.

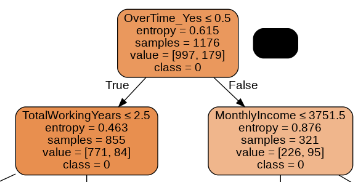
The upside of the Decision Tree Model is removing sections of the tree that are non-critical and redundant to classify instances. The model gives a simplified representation of the final classifier, which improves the predictive accuracy. The image below shows the best-pruned tree for our dataset. The best-pruned tree has a level of depth of 3 with the highest training accuracy of 72.4%. The depth level indicates how often the managers can make a decision before coming to a prediction.

Diagram

Description automatically generated

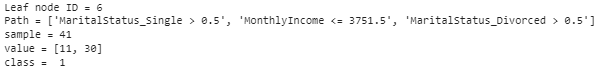
The data has yielded many different perspectives. Some were expected and some were not when we had originally given our personal insights. The data from the Logistic Regression has led us to believe that we should speak with employees and communicate to gather information if an employee meets the criteria and their probability to leave is over 50%. However, it is not enough to only know the riskiest variables but to find the interconnection between them. Having a list of conditions for employees can retrieve a better insight into the reason for employees leaving. With the help of the Classification Tree Model, we have a more straightforward visualization of what factors of employees would affect their probability of leaving the company.

Within the best prune tree, each box is called a node that contains information about the indicator for making a decision. For example, when an employee has not worked overtime, we can move to the node at the lower left corner and find that the probability of employees leaving the company is just 10% of their working years are less than 2.5 years. Otherwise, when a person is working overtime, HR has to look at the employee's monthly income factor. If the employee has a monthly salary less than $3751.5, the probability of them attriting is around 30%. The example illustrates how HR or manager would think about how likely the employee would leave based on the different signs from the employee’s background information.



Now, we understand how to apply the Decision Tree to the decision-making process. We are still curious about the highest probability of attrition with all things combined. The list below is the tree path for the best prune tree. We would like to select the top three paths with the highest predicted probability among the leaf nodes. The way to calculate the probability is by using the second number of the value row and dividing it by the sample size. The results are shown as follows.

| **Leaf Node ID** | **Sample Size** | **Value** | **Probability** |
| --- | --- | --- | --- |
| 1 | 47 | [26, 21] | 0.4468 |
| 2 | 23 | [22, 1] | 0.4348 |
| 3 | 625 | [588, 37] | 0.0592 |
| 4 | 160 | [125, 25] | 0.1563 |
| 5 | 72 | [44, 28] | 0.3889 |
| **6** | **41** | **[11, 30]** | **0.7317** |
| 7 | 126 | [116, 10] | 0.0794 |
| 8 | 82 | [55, 27] | 0.3293 |

*Leaf Node with the Highest Predicted Probability*

Once we know the probability for each leaf path, we will need to define an English Rule that ties back to the variables. The top three effective English Rule lies in leaf node 6, 1, and 2. The most effective English Rule has the highest predicted probability. Therefore, our conclusion for the Decision Tree is the following: if an employee's marital status is either single OR divorced AND monthly income is smaller than $3751.5, then the probability of an employee leaving the company is approximately 73%.

**Decision**

The HR Analysts have made some decisions in regard to which tops to present which may need to be looked over. If an employee has an overall probability of terminating their employment higher than 50% from the logical regression model, we would like to speak to those employees further and gather more information as to improvements that might be made to prevent separation. Looking deeper, we think the best approach would be to analyze the top 5 predictors. Several of those decisions are to gather more information as to why there is overtime and whether it would be more practical to hire additional personnel. Job Involvement is another predictor where we need to find out more information about whether there is too little or too much involvement. WorkLifeBalance could strongly be tied to overtime, but it would be beneficial to ask employees where it could be improved upon. Some of the work-life balance may be changing the business model to a more flexible hybrid model in which employees work from a combination of home and the office a certain number of days during the week. Job level could be tied to job involvement, sometimes being at a lower level in the company a person may not feel their talents are utilized while contrarily having a higher level in the company can make a person feel their talents are over-utilized. This would be worth gathering more information and exploring the sentiments of those with a higher probability score. Finally, looking at a position that does not require travel is highly variable and could also be worth looking at. Some of the non-travel positions could see the travel position having more flexibility, making them feel there is not a sense of WorkLifeBlanace, and would be worth exploring to gather more information.

**Advantages**

Many advantages can be derived from each insight. In general, the probability outcomes give us a clear overview of the likelihood of leaving the company for each employee. If we look at employees with a .5 or above the probability of separating from the company, we can gain a larger scope of the company. If we take the time to research more into overtime, we could not only reduce overtime we could add more diversity to the pool of current employees. If we were to research more into Job Involvement, this could lead to expanding more of the company's Diversity, Equality, and Inclusion (DEI) initiatives. If we look at worklifebalance variable, we may find that sometimes there are people with families who may not be able to make it to the office because of childcare issues, commuting issues, and other various life events. We could create a hybrid model of 2 days in the office and 3 days working from home and measure productivity levels to find if there is an increase over previous productivity while working inside the office. Job levels could give us a chance to create new mentor/mentee programs and give more experience for lower job levels and more perspective when making decisions for those at higher job levels. Finally, an advantage we could find for positions without non-Business travel would be to ask how we could improve the experience they have working in the office. This may lead us to increase more vacation time in which they could have time to travel and more family time, adding to work-life balance.

**Real-world Implementation**

1. **How we came out with the perspectives to address the problem**

With the former analysis using the logistics regression and classification tree, our actions will be designed according to the most important predictors that determine a person’s thoughts on whether to leave the company.

Among the top 5 predictors generated from the logistics regression analysis, we come out with 2 perspectives to address the problem: Work-life balance and an employee’s importance to the company. The work-life balance (rating), overtime, and business travel columns are the representations of an employee’s work-life balance level. While Job Involvement and Job Level could be representations of how important an employee’s role is to the company.

From the English rule we generated from the classification tree analysis, Marital Status is classified under the Work-Life Balance perspective while Monthly Income gives us a new perspective to think about this issue: Compensation. Please refer to the following table to understand how we distinguish the 2 perspectives more clearly.

|  | **Perspectives to Address the Problem** | | |
| --- | --- | --- | --- |
| **Importance of an Employee** | **Work-Life Balance** | **Compensation** |
| **Predictors** | Job Involvement | Work-Life Balance (Rating) | Monthly Income |
| Job Level | Overtime |  |
|  | Business Travel |  |
|  | Marital Status |  |

1. **the actions**

First, we should understand who to target so that the actions can be taken appropriately on the right people. One thing is to measure the job importance of an employee by looking at their job involvement and job level. If a person is a manager or above but has a high possibility of leaving the company, then the influence of this person’s leave could be large so it is more necessary that the leader or HR take some actions on the situation. Besides, in our predictions from logistics regression, our focus/target should be on the employees that have more than 50% but less than 80% possibility of leaving the company; this is because we also need to maintain the success rate of the actions being taken. If a person is over 80% likely to leave the company, there is not much flexibility or space for us to adjust our strategies to change to person’s mind.

After we narrow down the right targets, we suggest first properly conducting check-in meetings for the targets to meet with their supervisors/leaders in the team so that the team can directly adjust the routine or plans that might affect the person’s work-life balance if necessary (e.g. task arrangement). Besides, if sometimes overtime and business travels are necessary just because of the big environment (e.g. audit and tax busy season, stock market opening and close schedule, etc.), the company can also conduct well-being sessions to comfort employees’ mindsets more generally. The company could also send out surveys to get feedback on company policies if the overtime issue comes from the company culture that promotes a working atmosphere that’s too challenging.

Another method to decrease employees’ attrition rate is through compensation as monthly income is also a factor from the English rule of classification tree analysis. One option is that the leaders of the team or HR could consider improving employees’ monthly salaries reasonably based on the industrial average. The other option, however, is to increase the share of stocks that the employees are holding, because the stock might be a more important predictor than salary according to the results from logistic regression analysis. This also makes sense in reality because as the company stocks are growing, the employee will also gain responsibility and a sense of belonging that reduces their attrition probabilities.