**CIS 8693 DATA ANALYTICS FOR DI**

**Project Group #6**

**EMPLOYEE ATTRITION ANALYSIS**

**Anindita Sarmah** [**asarmah1@student.gsu.edu**](mailto:asarmah1@student.gsu.edu)

Panther ID: 002733978

**Ariba Surani** [**fariba1@student.gsu.edu**](mailto:fariba1@student.gsu.edu)

Panther ID- 002676489

**Mayuri Jadhav** [**mjadhav1@student.gsu.edu**](mailto:mjadhav1@student.gsu.edu)

Panther ID- 002732643

**Sanika Dhawale** [**sdhawale1@student.gsu.edu**](mailto:sdhawale1@student.gsu.edu)

Panther ID- 002733251

**Guided By: Prof. Balasubramaniam Ramesh**

**Employee Attrition Analysis**

**Project Summary**

It is vital for organizations to analyze the employee attrition rate and the factors of attrition as it helps companies to be prepared for future employee loss. Analysis helps companies prevent valuable employees from leaving and find the reasons why their employees are leaving.

We analyzed the attrition rate from a fictional dataset - ‘IBM HR Analytics Employee Attrition & Performance’ created by IBM data scientists sourced from Kaggle. Our dataset contains 35 columns and 1470 records. There are no missing values in the dataset.

**Dataset:** <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

After the data collection step, we explored and cleaned the data. We used Azure machine learning to run models like Neural network, Logistic Regression, Decision tree and Boosted decision tree on our dataset and used Heuristic lab to run the Generic algorithm model.

The document will look in more detail at all these steps and a conclusion will be made.

**Data Exploration and Preparation**

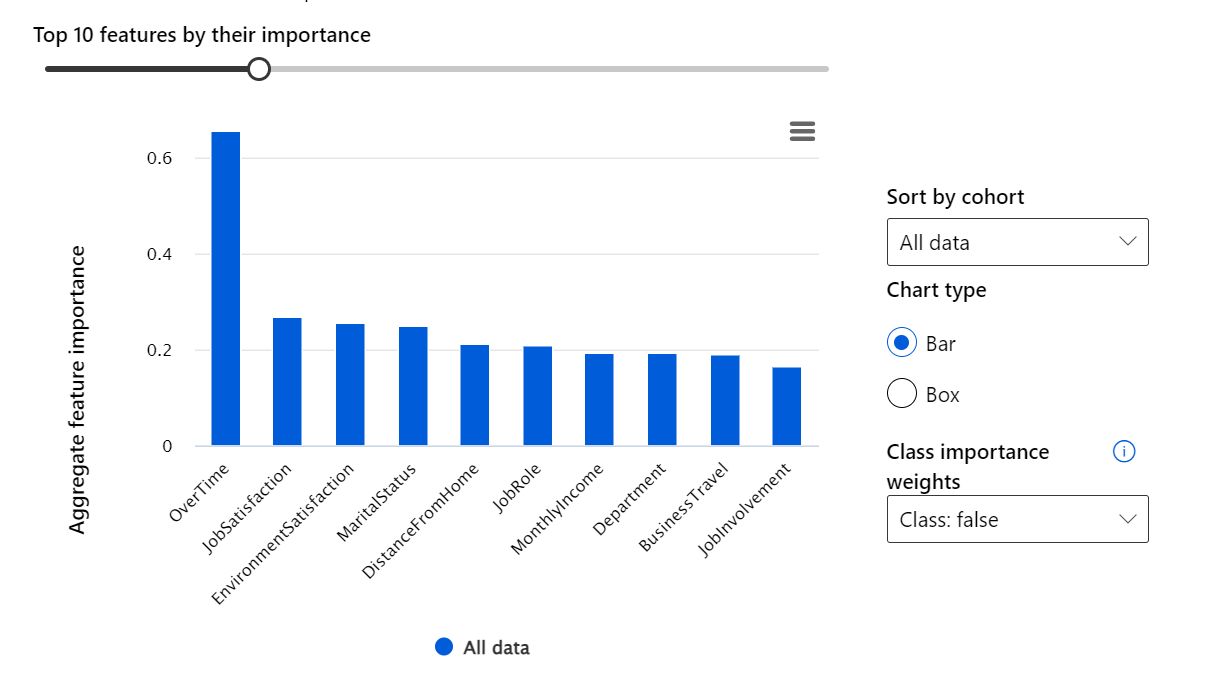
In order to clean the data, we deleted Over18 and Education field as these were creating very little variance with the target variable. We refined the columns Attrition and Over Time where we changed the Yes/No String inputs to binary inputs 1/0.

**Tools Used:**

* Azure ML
* Heuristic Lab
* Tableau

**Dependent and independent variables**

The dependent variable in our data set is the employee attrition rate -’Attrition’ and the independent variables are BusinessTravel, DailyRate, Department, EmployeeCount, Employee Number, Gender, Hourly Rate, Job involved, Job level, Job role, Job Satisfaction, MaritalStatus , MonthlyIncome, age, relationship status and the years spent on this and other job roles. The top 10 important features are displayed below:



**1.Logistic Regression**

The first model that we used was logistic regression. It is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, when you are trying to determine if a new sample fits best into a category.

Graphical user interface, chart, line chart

Description automatically generated

As shown in the above screenshot, the accuracy of Logistic Regression is 0.891 which is very high, the precision is 0.872, the number of predicted false positive which were false are 395 which is significantly very high.

**2.Decision Tree**

Next, we used Decision trees which build classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor is called root node. Decision trees can handle both categorical and numerical data. Decision trees are simple to understand even for someone who might not be familiar with data analytics.

Graphical user interface, chart, line chart

Description automatically generated

As shown in the screenshot above the accuracy of Decision tree is 0.816 which is very high., the precision is 0.457, The number of predicted false positive which were actually false are 335 which is significantly very high .

**3. Boosted Decision Tree**

We went deep and tested the boosted decision tree model. *Boosting* means that each tree is dependent on prior trees. The algorithm is learned by fitting the residual of the trees that preceded it. Thus, boosting in a decision tree ensemble tends to improve accuracy with some small risk of less coverage and as predicted our accuracy improved.

Graphical user interface, chart, line chart

Description automatically generated

The accuracy of Decision tree is 0.859 which is higher than that of a normal decision tree, the precision is 0.703 which is again higher than a standard decision tree, The number of predicted false positive which were actually false are 353 which is significantly high.

**4.Neural Network**

Next in our analysis was Neural Network. A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It produces the highest level of accuracy in comparison to the above models but low explainability and it's complicated to explain.

Graphical user interface, chart, line chart

Description automatically generated

As shown in the screenshot above the accuracy of Decision tree is 0.873 which is very high., the precision is 0.756, The number of predicted false positive which were actually false are 354 which is significantly very high.

**5.Genetic Algorithm**

Lastly, we used Genetic Algorithms to test the accuracy of our dataset. The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. The genetic algorithm repeatedly modifies a population of individual solutions. Even though GA produces a highly accurate result it is very complex to understand it.

Graphical user interface, application, table, Excel

Description automatically generated

The R2 is 0.045, the mean absolute error is 2.57. Surprisingly GA did not propose a very high accurate result for our dataset

**Conclusion**

After running the dataset through the various data analysis models like Logistic Regression, Decision Tree, Boosted Decision Tree, Neural Network and Genetical Algorithm, we observed the below options:

* Logistic Regression model has the highest accuracy out of all
* Neural Network has the next highest accuracy, which is very close to Logistic Regression
* But Neural Network has very low explainability, whereas Logistic Regression has decent explainability
* Decision tree can help achieve better explainability but there is a trade-off with accuracy
* Genetic Algorithm has the least accuracy

Hence Logistic Regression is the best suitable model for Employee Attribution problem

**Recommendations and Insights**

So, through the Logistic Regression analysis, the most important features affecting the target variable are Overtime, Job Satisfaction, Environment Satisfaction and Marital Status from which we can draw the following recommendations and insights which can help the company to reduce the employee attrition rate:

* Increase allowance for employees to work overtime and make sure those who are working overtime are compensated fairly
* Company should remember to appreciate employee's work using intrinsic and monetary factors
* Fairly pay employee in same job level and job involvement with almost same income and stocks
* Employer should treat everyone fairly and do not discriminate on basis of marital status