**Milestone-1**

* 1. **Introduction**

This project will focus on analyzing the data collected by the HR department and on building a model that will predict whether an employee will leave or stay in the company. For this project we will try 3 machine learning algorithms. These are Random Forest, SVC, and Logistic Regression. We will fine tune our model using bagging and boosting as well as finding the best parameters for our model using Randomized Search CV. Finally, we will evaluate all the models and choose the best performing one.

The company will also benefit by identifying factors that affect their decision to leave. It is time-consuming and expensive to find, interview, and hire new employees; thus, increasing employee retention will be advantageous to the company.

# **Business Objective**

A fictional French-based alternative energy vehicle manufacturer, Salifort Motors is concerned about employee turnover rate. The company strives to create corporate culture that supports employee success and professional development. They collected data from employees and is need of a data analytics professional that will provide data-driven suggestions based on the data.

They have the following question:

* What factors affect employee retention?
* How to increase employee retention?
* Build a model to see which employee is likely to leave.
* Analyze the attributes of those employee to improve satisfaction.
  1. **Content**

The dataset used in this case study is located on Kaggle and it contains 15,000 rows and 10 columns for the variables listed below. It has 8 numerical and 2 categorical variables.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThis dataset contains columns: Satisfactory Level, Number of Project, Average Monthly Hours, Time Spend Company, Promotion Last 5 Years, Department, Salary.

**1.4 Challenges**

* Handle missing values
* Detect and clean outliers
* Plot visuals to find correlations between variables
* Check skewness of data
* Feature selection
* Convert categorical values to numerical values using encoding
* Split data into test and train sets
* Train model
* Validate model using cross validation
* Measure accuracy of model using different metrics
* Improve model performance using bagging or boosting
* Prepare a dashboard to keep management informed about the situation of employees in the company

**Milestone-2**

* 1. **Prediction Techniques:**

To predict if an employee will leave the company or not, we will use different CART algorithms and determine which one gives us a better performance. By exploring the different parameters, we can have a better understanding of which factors affect our model. Finally, a BI dashboard, will help company officials informed about the situation.

* 1. **Data Information:**

A screenshot of a computer program

Description automatically generatedData has 15000 rows and 10 columns. All columns have some null values which will be handle in future steps. 8 columns have float values and 2 have categorical values.\

Taking a closer look at the categorical values. There are 10 departments in the company, sales and technical have the highest employees while management has the least employees. The salary column has three ranges, low, medium, and high.

* 1. A screenshot of a computer

     Description automatically generated**Basic Statistics**

**A screenshot of a computer

Description automatically generated**

* As we can see more than 50% of employees have a satisfaction level of over 0.6.
* Average time spent in the company is higher than the 50th percentile, this means that there are newer employees.

**Milestone – 3**

**3.1 Correlation**

A colorful squares with black text

Description automatically generated with medium confidence

The matrix shows most attributes have low correlation between each other. Only ‘last\_evaluatiion’, ‘num\_project’ and ‘average\_monthly\_hours’ worked has slightly positive correlation. On the other hand, work accident and employee retention have some negative retention. This means that employees tend to leave if some accident occurs at the workplace.

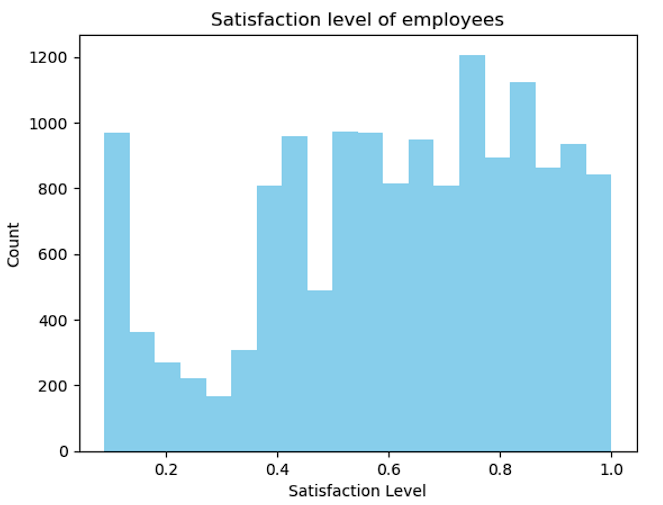
**3.2 Handling missing values**

The data contains very small percentage of missing values. For numerical columns I used the mean to fill the missing values. For the categorical attributes I removed the rows.

A screenshot of a computer code

Description automatically generated

**3.3 Visualizations**

First, we look at the distribution of satisfaction level of employees.

A graph of a number of employees with satisfaction

Description automatically generatedThis shows that most employees have satisfaction level of over 50%. Now we will look at people who has more than 50% satisfaction but left the company. Values at 0 means that employees stayed and 1 means that they left.

Next, we look at employee status in different departments.

A graph with numbers and text

Description automatically generated

The most employes left from sales department, but the ratio of left: stay is lowest in HR department. This means people from HR are most likely to leave.

Finally, we look at the time spend by each employee at the company.

A graph with blue squares

Description automatically generated

This shows that most employees have 2-6 years of experience in the company. We can look at the employee who stayed more than 6 years but still left. Surprisingly, no employee with more than 6+ years of job time left the company.

**3.4 Correlation Plots**

**A group of blue squares

Description automatically generated**

**3.5 Outlier Detection**

I removed outliers in the 'satisfaction\_level' column of the HR dataset. The Z-score method is employed for outlier detection, allowing us to identify data points that deviate significantly from the mean. Z-scores are calculated for each data point, indicating how many standard deviations a value is from the mean. Outliers are defined as data points with an absolute Z-score greater than 3. The final data has shape (14941, 10). The histogram below shows the distribution in ‘satisfaction\_level’.

A diagram of a box diagram

Description automatically generated

**Milestone – 4**

**4.1 Skewness of data**

A number of numbers on a white background

Description automatically generated

This shows most columns has bell shaped curve. Only last\_evaluation has some right skewness.

A graph of a person with a black line

Description automatically generatedA graph of a normal distribution

Description automatically generated

A graph with blue lines and black text

Description automatically generatedA graph of a normal distribution

Description automatically generated

We will transform the last evaluation to get a more balanced data. For this we can use boxcox method. One challenge that I faced while transforming the data is that data contains 0 values. To handle this, I added 1 to all the data and used the transformation technique. The results are as follows:

A graph of a normal distribution

Description automatically generated

**4.2 Encoding**

Two encoding techniques, namely Label Encoder and Ordinal Encoder, are applied to the 'Department' and 'salary' columns, respectively. The Ordinal Encoder ensures that hierarchy is maintained in the data. For departments, we can use Label Encoder as we don’t need any order there. The encoded data:

A screenshot of a computer

Description automatically generated

**4.3 Objective solved (1)**

* What factors affect employee retention?

We can see that time spend in the company plays a significant role in employee retention. Employees with more than 6 years at the company has a very low possibility to leave. People in HR department are likely to leave, compared to other departments. This may be due to some internal problems not reflected in the data. All other variables have some effect in retention as well.

**4.4 Modeling**

To start our modeling process, we will divide the dataset into train and test sets. We will use the ‘left’ column as our dependent variable and all the other columns as independent variables. We will use random state as 40 and test size as 0.3 to use 30% as test data and rest as train. Random state ensure we get the same test/train data every time we run the code.

**Y- column**

A screenshot of a computer code

Description automatically generated

**X- columns**

A table with numbers and a few words

Description automatically generated with medium confidence

**Milestone- 5**

**5.1 Model Selection**

I will be testing out 3 different models to see which performs the best:

1. Logistic Regression: model which uses the probability of a discrete outcome given an input variable. Uses Sigmoid function to make predictions.
2. Support Vector Classifier: is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes.
3. Random Forests: combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

**5.2 Model Evaluation**

To evaluate our models, we will be using 3 different metrics. These are classification report, accuracy, and confusion matrix.

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated

A screenshot of a screen

Description automatically generated

All the F-1 scores are above 0.85 which means our data is well balanced. As we can identify from the scores, Random Forest Classifier has the best performance. The confusion matrix shows that we only have 8 FP and 35 FN We will select this model and perform cross-validation. The result of cross-validation shows a higher accuracy for our model.



**5.4 ROC curve**

The ROC curve plots the True Positive Rate (Sensitivity or Recall) against the False Positive Rate at different classification thresholds. A curve that approaches the top-left corner of the plot indicates a better-performing model. In our model, AUC-ROC of 0.98, we have a very high AUC-ROC value. A value of 0.98 is considered outstanding, and it indicates that our model is performing exceptionally well in terms of ranking and classifying data.

A graph with a line and a random line

Description automatically generated with medium confidence

**5.5 Objective Solved**

* How to increase employee retention?

Fix internal problems in HR department, where employee retention is the lowest. This may include better incentives or improved work environment.

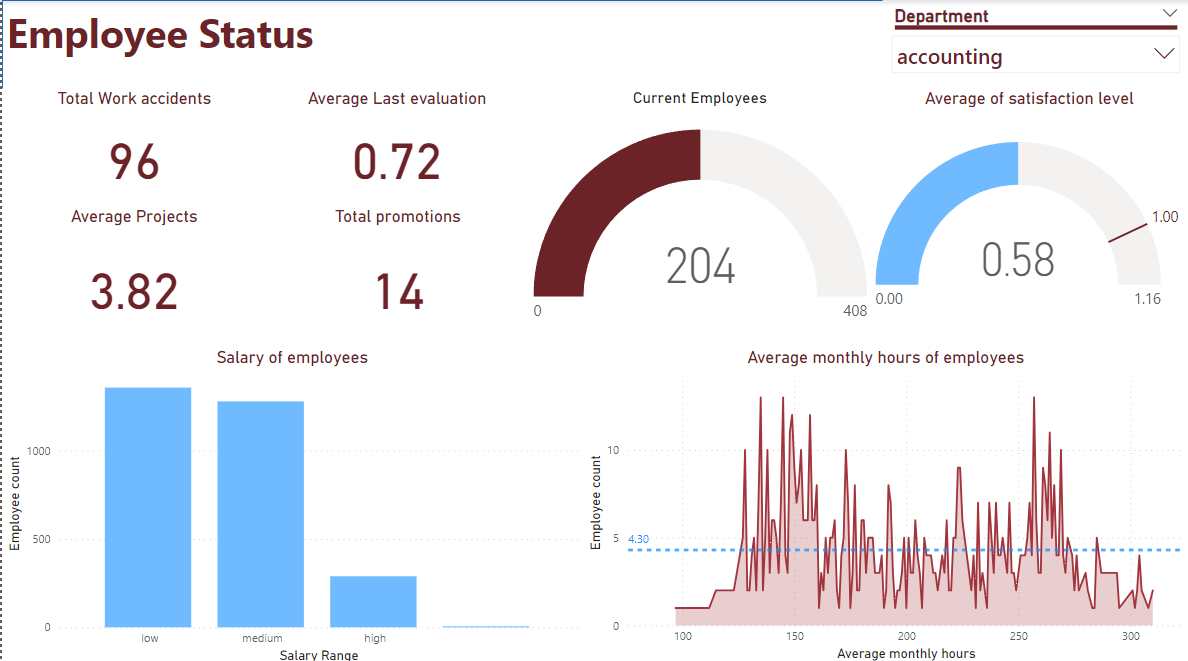
* Build a model to see which employee is likely to leave.

Random forest classifier is the final model to test whether an employee is likely to leave. Model has an overall accuracy of 99.18%.

* -Analyze the attributes of those employee to improve satisfaction.

Improve employee satisfaction by giving better incentives. Analysis showed that employee with low salary is likely to leave. This mean that they feel they are under paid and have better opportunities somewhere else.

**5.6 Dashboard**



**Milestone – 6**

Did I have a clear understanding of the project and define the goals and objectives?

* Yes, my goals were clear in the start, and I worked to complete them.

How did I approach handling and working with the dataset?

* Dataset was available on Kaggle and there were no issues with importing the file

What steps did I take to complete the project?

* Define business objectives, conduct exploratory analysis, build model, and evaluate it.

Why did I choose a specific model?

* Random Forest had the highest accuracy and is a well-known model to handle classification and Regression tasks.

What guided my decision-making process?

* The objectives I set out at the start of the project.

How the introduced model in my project can help others interested in or working I the area related to the subject of my project?

* They can apply this model to determine the retention of employees

How did I analyze and present the project's results? Was there anything I could have done differently?

* I used cross-validation and ROC curve to present the results.

Could I predict the (short/medium/long term) future of the subject under study in my project using the available data in the data set I chose?

* Yes, I can predict medium term future based on my model.

What are the most important analytical tools or modeling techniques presented in this course that I applied in my project?

* Classification matrix, Scatter plots, Histograms and Random Forest classifier.

Why did I find them useful?

* Yes, they helped me complete my project.

What challenges did I face during the project, from cleaning to modeling my data set and how did I overcome them?

* Skewness of data – Used Boxcox to transform them
* Encoding – Used encoding to handle categorical values
* Outlier detection- using Z-scores, determined outliers and removed them from my data
* Preparing a dashboard to keep officials informed about company status

What new skills or knowledge did I maintain from this project, and how did theys impact my work or personal growth

**References:**

Dataset: [*https://www.kaggle.com/datasets/mfaisalqureshi/hr-analytics-and-job-prediction/data*](https://www.kaggle.com/datasets/mfaisalqureshi/hr-analytics-and-job-prediction/data)