Human Resources Analytics

## Problem Definition and Planning:

Description:

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization. The goal of the HR Analytics is to provide an organization with insights for effectively managing employees so that business goals can be reached quickly and efficiently. Here we are trying to get some insights about why valuable employees are leaving prematurely and who will leave next so that company can either take appropriate measures to retain them or they can plan so that company’s performance will not get affected.

* + - Problem Statement:
* Why Valuable employees are leaving prematurely?
* Which employees will leave next?
  + - Project deliverables:
* What makes an employee valuable?
* Who all are valuable employees?
* Define premature exit
* Categorize employees based on their major reason of exit
* Based on the above categories suggest measures to retain valuable employees
* Select an effective model to predict premature exit of employees

* + - Generate Success factors:
* Accuracy measures (Laplace and M estimate)
* ROC: Graphical plot that illustrates the performance of binary classifier as its discrimination threshold is varied
* Confusion matrix (precision and recall): **Recall**: Total number of documents retrieved by the search divided by total number of existing relevant documents.

**Precision**: Total number of documents retrieved by the search divided by total number of elements that belong to positive class

* + - Resources:
      1. **Software tools:**
* Asana- Project Management Tool
* GitHub- Version Control Tool
* R Studio and R
* Tableau and Plotly for visualization
  + - 1. **Project Members:**
* Divya Vasireddy
* Naveen Sampath
* Nivedita Merla
* Snehal Gawas

## Data preparation:

* + - Data Source:

Kaggle Dataset: <https://www.kaggle.com/ludobenistant/hr-analytics>

* + - Summarize data:

Number of observations = 15,000

Number of variables = 10

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| satisfaction\_level | Level of satisfaction | Numeric |
| last\_evaluation | Evaluation of employee performance | Numeric |
| number\_project | Number of projects completed while at work | Numeric |
| average\_montly\_hours | Average monthly hours at workplace | Numeric |
| time\_spend\_company | Number of years spent in the company | Numeric |
| Work\_accident | Whether the employee had a workplace accident | Numeric |
| left | Whether the employee left the workplace or not | Numeric |
| promotion\_last\_5years | Whether the employee was promoted in the last five years | Numeric |
| sales | Department in which they work for | String |
| salary | Relative level of salary | Numeric |

* + - Transform data:
* We need to assign appropriate class for each varaible. *number\_project and time\_spend\_company* must be integrs. *Work\_accident, left, promotion\_last\_5year and sales* must be nominal. *salary* must be ordinal
* Rename variables to meaningful names
* Add a unique identifier for each employee
* Summarizing the data using appropriate statistical measures and plots
  + - Segment:
* Divide the data into train and test sets based on *left* variable. Meaning, the proportion of 1’s and 0’s in the train and test should be same.
* Training set will be used for building classification model and test set will be used to validating the results.

## Analysis:

* + - Summarize Data:

Summarization is an important step in the data analysis process as it reduces data and provides interpretation of without losing any data.

In this phase, we our focus will be on generating summaries for the different departments in the company like Accounting, Sales, Hr, Marketing, IT etc. plus summaries based on different factors like time spend in company by an employee, salary and working hours.

* + - Exploring Relationships between attributes:

This is an important phase selection of attributes affect the accuracy of the model. In this phase, we will try to understand the relationships between the attributes. If we found any correlated attributes, then we might have to remove them from the model.

* + - Regularization:

We wish to develop a simple model which will be a sweet spot between flexible and conservative model. To achieve this objective regularization is an important step. Penalty will to added for complexity and it will force the model to stay simple.

We are planning to use Lasso which is the combines both strategies Selection (Selecting best features) and Shrinkage (Reduce influence of each feature in case if it turns out to be noisy).

* + - Grouping Data:

We will use clustering method to group employees who have already left to understand similar attributes that are affecting their exit.

* + - Identifying Patterns, Trends and Non-Trivial Facts:

Based on above groups we hope to find some trend in the data which can highlight the reasons behind the employee. This will also help in identifying anomalies in the data.

* + - Predicting Model:

We will be exploring below listed methods to achieve our objective. The best model will be selected based on the accuracy of the model.

* + - 1. **Logistic Regression:**

Logistic regression is a regression model used to predict categorical dependent variable. In our case, we will try to predict whether employee will leave (1) or not (0).

* + - 1. **Decision Trees:**

Decision tree is a tree like model of decisions. It represents decisions, possible consequences and their chance. We have chosen this model because it is simple to understand and interpret.

* + - 1. **Random Forest:**

It is an ensemble learning method for classification or regression. It constructs multitude of decision trees at training time and output the class that is mode of the classes of individual trees. We have selected this method because it corrects overfitting habit of decision trees.

* + - 1. **KNN – (k – nearest neighbours):**

It is a non-parametric method for classification. An object is classified by majority vote of its neighbours and object gets assigned to the most common class among its k nearest neighbours.

* + - 1. **Gradient Boosting:**

It is one of the technique used to solve classification problem. It produces a prediction model in stage-wise fashion in the form of weak prediction models and generalizing them by allowing optimization of an arbitrary differentiable loss function.

* + - 1. **Support Vector Machine:**

SVM training algorithm builds a model that assigns new examples to one of the category, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.