**Feature Engineering Techniques**

Feature Engineering refers to the process of designing artificial features into an algorithm. These artificial features are then used by that algorithm to improve its performance, or in other words reap better results.

* **Features Selection:**

|  |  |
| --- | --- |
| **Filter Method** | **Chi Square & Mutual Info method** |
| * Measure the relevance of features by their correlation with dependent variable, doesn’t train any model * Relies solely on properties of the data, thus independent on any algorithm. * Faster and less computationally expensive * Avoids overfitting * We use Pearson Correlation Technique   Measures the strength of relationships between only two variables, without taking into consideration the fact that both these variables may be influenced by a third variable | * Measure the usefulness of a subset of feature by training a model on it * uses the inductive algorithm to estimate the value of a given subset * Slower and computationally expensive * More prone to overfitting * We use Sequential backward selection   Removes (backward selection) features to form a feature subset in a greedy fashion  At each stage, this estimator chooses the best feature to remove based on the cross-validation score of the estimator "Linear Regression" |

|  |  |
| --- | --- |
| **Feature ranking** |  |
| * Feature ranking refers ordering the features from best to worst. * Using the Random Forest's feature importance attribute, we rank the features. * Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the features |  |

**Class imbalance treatment**

* Handling class imbalance is important as we need to train a machine learning model that generalizes well for all possible classes
* To get better class clusters, Tomek links are applied to oversampled minority class samples done by SMOTE (Synthetic Minority Over-sampling Technique). Tomek links are pairs of instances of opposite classes who are their own nearest neighbours

**Overview**

* **Objective**

To Analyze and Build models for Predicting Customer Rating using machine learning techniques

There are several reasons for employee attrition from an organization, which include better pay and job opportunities outside the organization, low Pay, or no recognition of employees in the current organization, improper work life balance, impolite behavior of managers and peers leading to ineffective team management, poor quality of work life, poor working conditions or even death of employees when at job, etc. Employees are the backbone of an organization. Hence, employee attrition can affect the organization adversely.

So, employee attrition analysis becomes a highly prioritized task for every organization

* **Methodology**
* Get maximum insights from a data set
* Uncover Underlying structure
* Extract important features from the data set
* Detect Outliers and anomalies (If any)
* Train a Machine Learning model for predicting Employee Attrition prediction with Python
* Validation of Predicted Model
* Visualization of results with Graphical representations

**Methodology**

The two machine learning techniques used

**Decision Tree Technique**

* Non-Parametric Supervised learning model, simple decision rules inferred from the data features.
* A decision tree (also a directional graph) is a series of nodes with Decision Nodes and Leaf Nodes

**Logisticl Regression:**

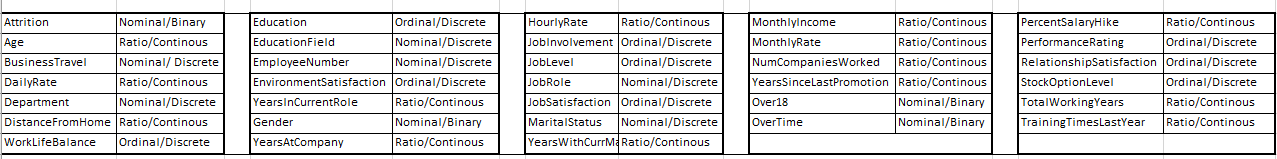
* Logistic regression - Parametric Supervised learning model based on probability
* Used to calculate or predict the probability of a binary (yes/no) event occurring.

**Dataset**

* **Size of the dataset**

33 Features, 1470 Records/Data Points

* **Variable type**



* **Data Distribution and handling imbalanced data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Distribution:**  Classification of Attrition Data:   * High - 659 * Low - 341 |  | | **Handling Imbalance Data:**   * Imbalanced dataset with a greater number of 'No' compared to 'Yes' * Handling the Dataset Imbalance Using Hybridization: SMOTE + Tomek inks * To get better class clusters, Tomek links (Pairs of instances of opposite classes who are their own nearest neighbours) are applied to oversampled minority class samples done by SMOTE (Synthetic Minority Over-sampling Technique) |
| **Feature Wrangling** | | | **EDA outcomes and discussion** | | |
| 1. Data exploration - analysis of dataset by visualizing representations of data 2. Dealing with missing values  * Check Missing values, Data types, wrong entries * Visualizing the missing values  1. Reshaping data  * Normalization/Standardization * PCA Dimensionality reduction * Encoding the data  1. Filtering data - Removal of Unwanted columns not required for Prediction Analysis | | | * Visualization of Data using the Heatmap – this gives the insights of the data from Correlation perspective (Positive Correlation, Negative Correlation) * Finding Bi relations and Multiple relationships with Visualization methods   **Pic represent the relations between Attrition with Age, Daily Rate and Monthly Income** | | |

**Results**

### **Table for the evaluation metric for each ML technique used**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature Selection1 – Filter Method:** | | | **Feature Selection2 – Chi-Square and Mutual Info** | | |
|  | | |  | | |
| **Plot Curve – Logistic Regression** | **Plot Curve – Decision Tree** |  | **Plot Curve – Logistic Regression** | **Plot Curve – Decision Tree** |  |
| **Conclusion** | | |  | | |
| From testing both models with the test set, we concluded that the performance of each factor varies with respect to both method   * Decision Tree has the highest accuracy, meaning it guesses correctly 81% and 79% of the predictions for both feature selection methods. * Considering F1-score, we can find that for Decision Tree and Logistic Regression, F1-score (which gives a balance between precision and recall) is same while predicting 'Yes' and the F1-score is higher for Logistic Regression while predicting 'No' * ROC curve is also a good measure to choose the best model. Larger the area under the curve (AUC), the better is the model. Here we have larger area for Logistic Regression.   **Thus, considering all the factors, we conclude that 'Logistic Regression' is the better model** | | | | | |