Shobha

HR ANALYTIC PROJECT



**Abstract**

Employee attrition can become a serious issue because of the impacts on the organization’s competitive advantage. It can become costly for an organization. The cost of employee attrition would be the cost related to the human resources life cycle, lost knowledge, employee morale, and organizational culture. This study aimed to analyse employee attrition using logistic regression, Random Forest Classification, Decision Tree Classifier and SVC. The result obtained can be used by the management to understand what modifications they should perform to the workplace to get most of their workers to stay. We use python for data integration, exploratory data analysis, data preparation, logistic regression, Random Forest Classification, Decision Tree Classifier and SVC model evaluation, and visualization. The study has five steps: (1) data collection and business understanding, (2) data pre-processing, (3) exploratory data analysis, (4) model selection and training, and (5) test and evaluation of the model. The result of the study found eleven variables as key driving factors for employee attrition.

**Introduction:**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

Under HR analytics project we are analysing that what are the important factors which influenced attrition rate in an organization which results high cost to an organization. High attrition become a problem to any business. Our Study is based on 35 factors which can be a reason for high employee attrition.

The input dataset is an Excel file with information about 1470 employees. For each employee, in addition to whether the employee left or not (attrition), there are features such as age, employee role, daily rate, job satisfaction, years at the company, years in current role, etc.

The steps we will go through are:

Data Pre-processing: pre-process the data to suit them with the analysis method.

The pre-processing may involve cleaning up the data, transforming the data, or creating new variables that may bring useful information for the analysis steps.

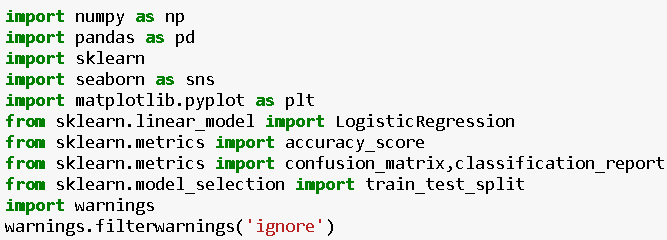
Exploratory Data Analysis (EDA): this step creates textual and visual summaries of the dataset that highlight some characteristics of the data.

Model Selection and Training, Test and Evaluate the Model: evaluate the performance of the proposed model

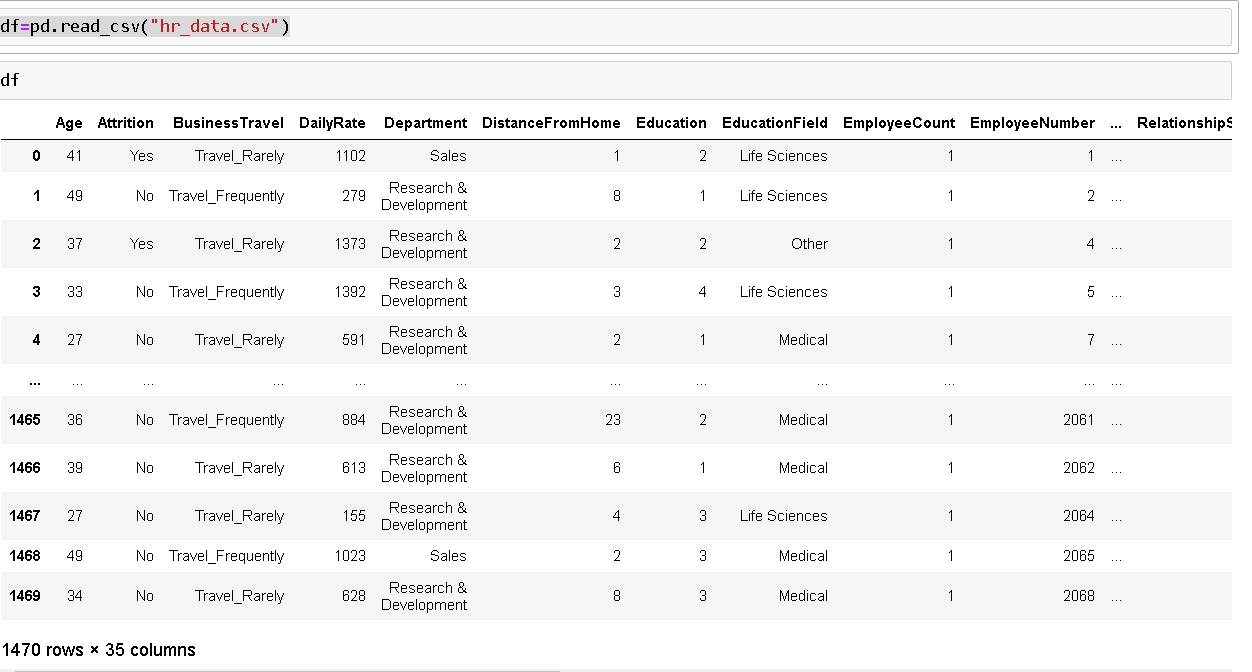
Data pre-processing

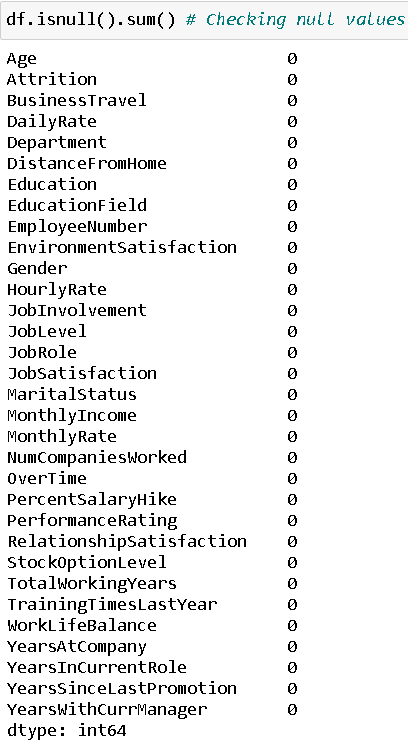
Getting the system ready and loading the data

We will be using Python for this course along with the below-listed libraries.



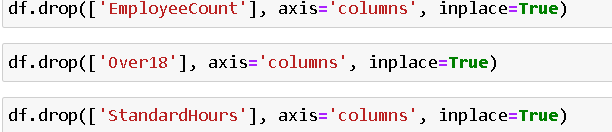
**Reading Data**



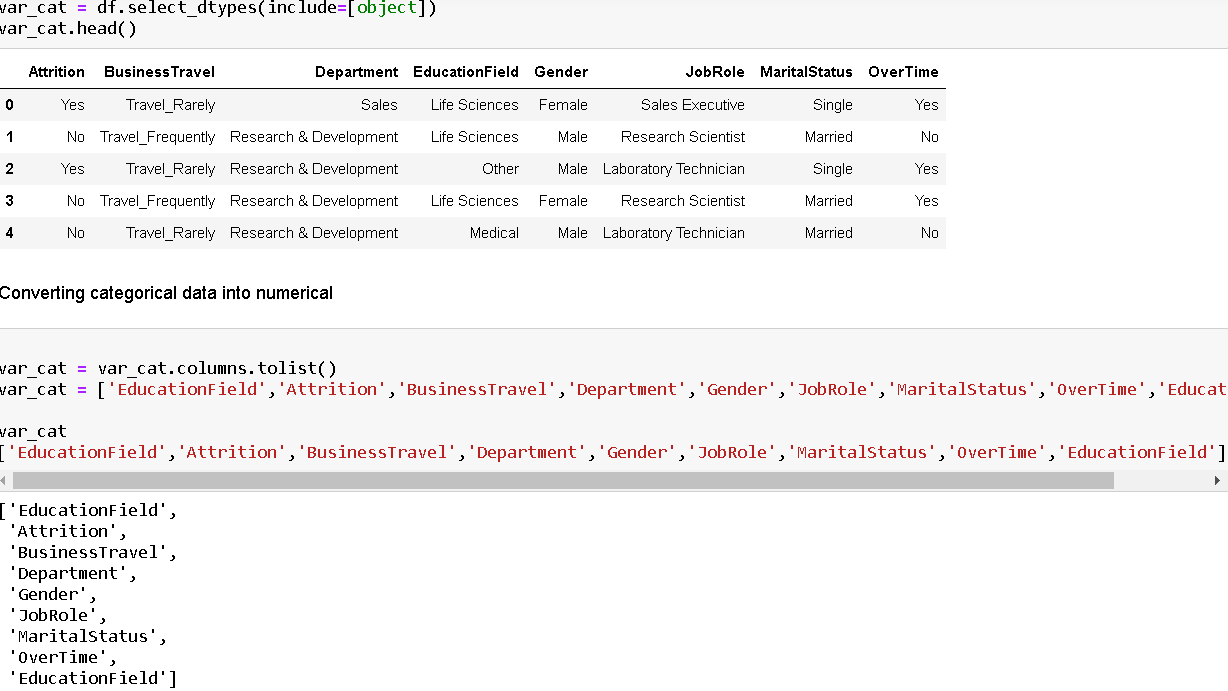


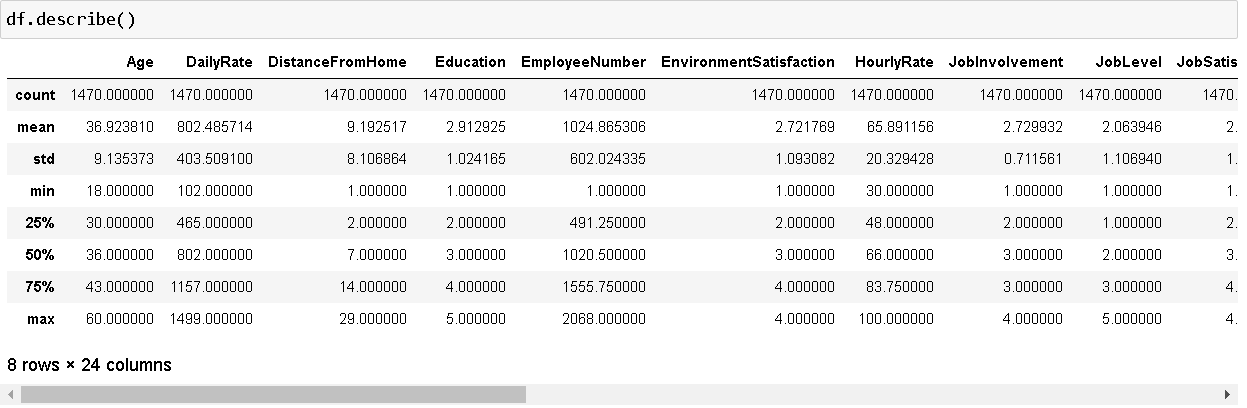
**Good there is no null values in the given dataset.**

**We are removing Employee count, Over18 and Standard Hours as all of the columns have same values and hence not contributing much to predict in model accuracy.**

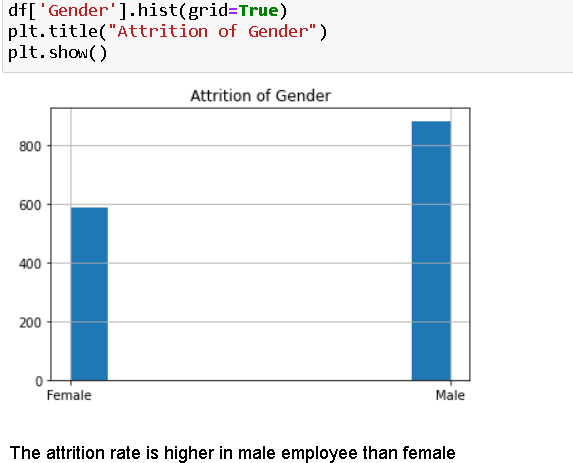
****

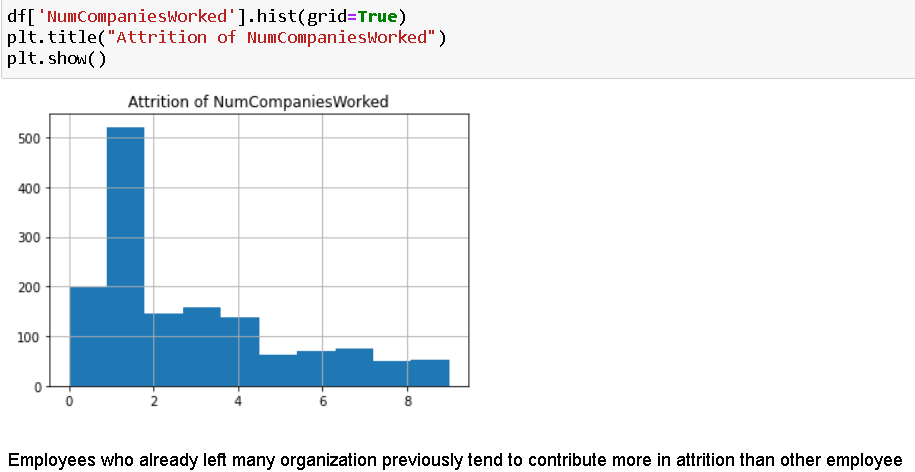
**Converting string values**

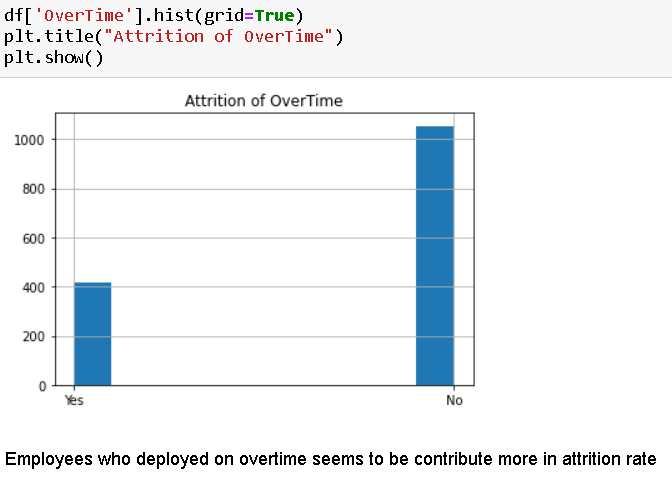
****

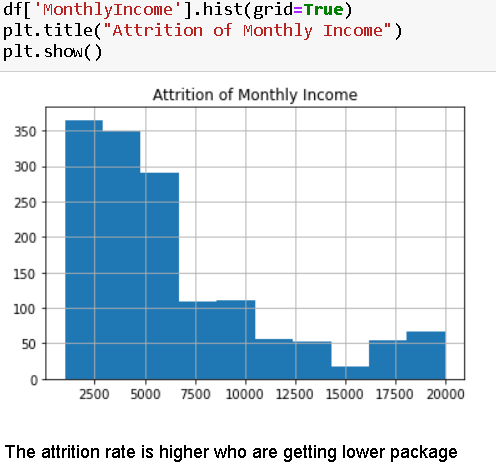
****As per above statistics description there is no null value in given dataset. Most of the outliers falls in Total working year, training times last year, Years at company, years in current role, years since last promotion, years with current manager and education field. Standard deviation is on higher side in some of the columns. In some of the columns there is difference in mean and 50% and mean and standard deviation so the data is again skewed in columns.

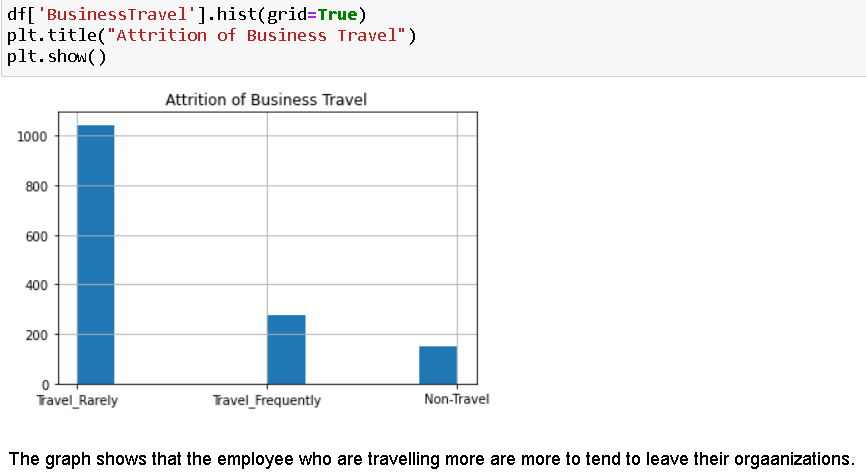
**Data Visualization:**

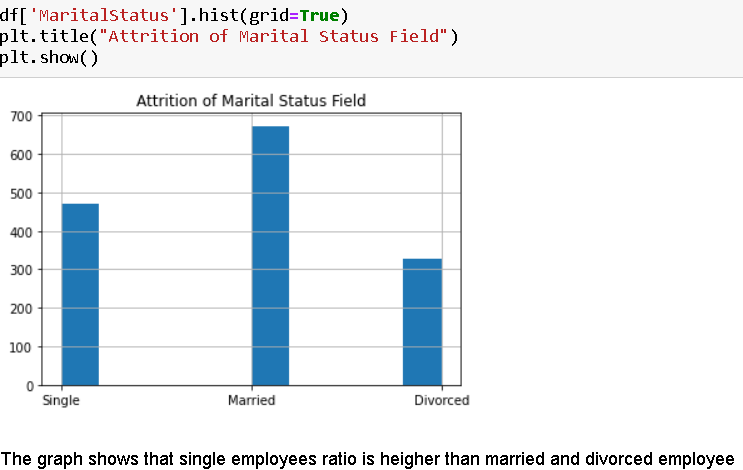
****

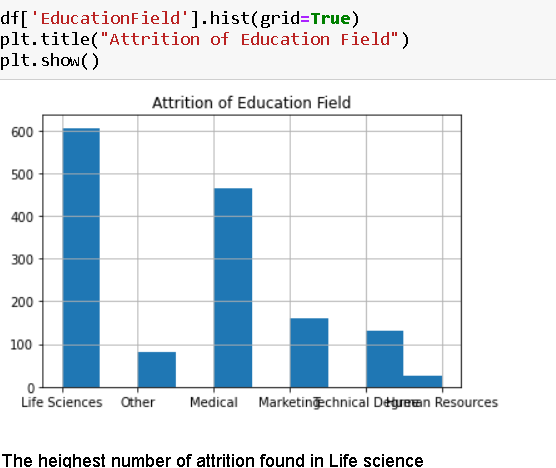
****

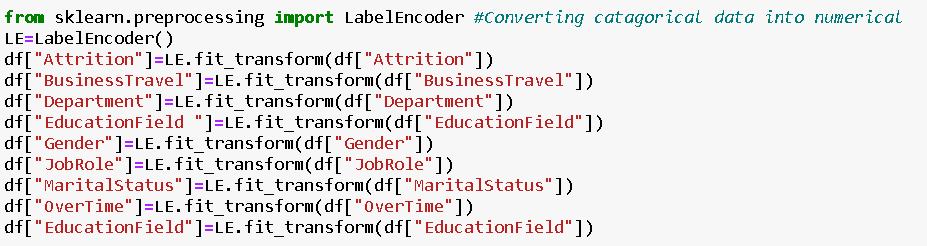
****

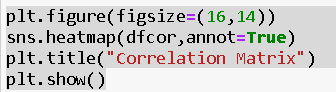
****

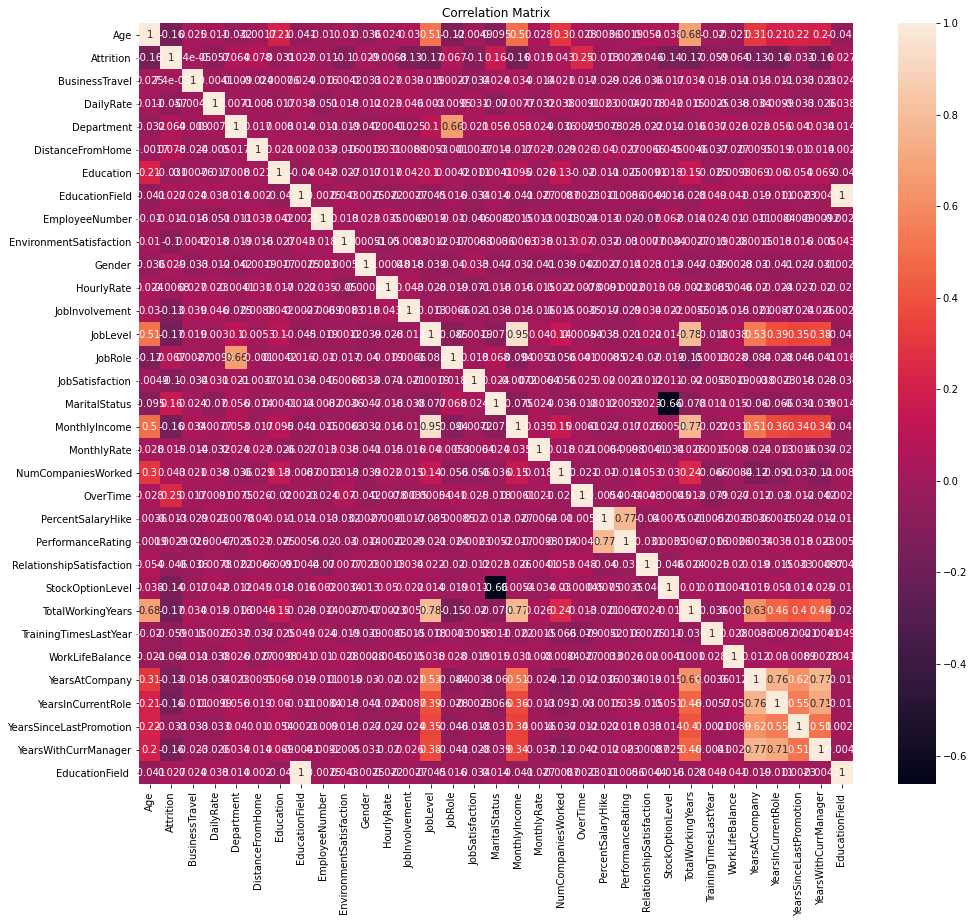
****

****

****

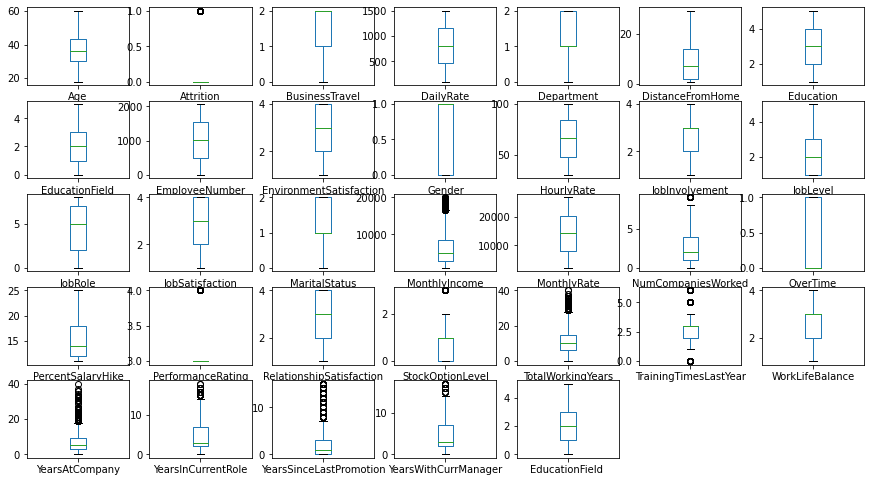
****

****

****

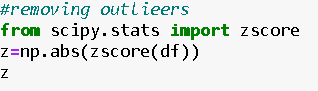
As per above graphics some of the variables are moderately correlated with Attrition and some of the variables are negatively correlated with attrition.

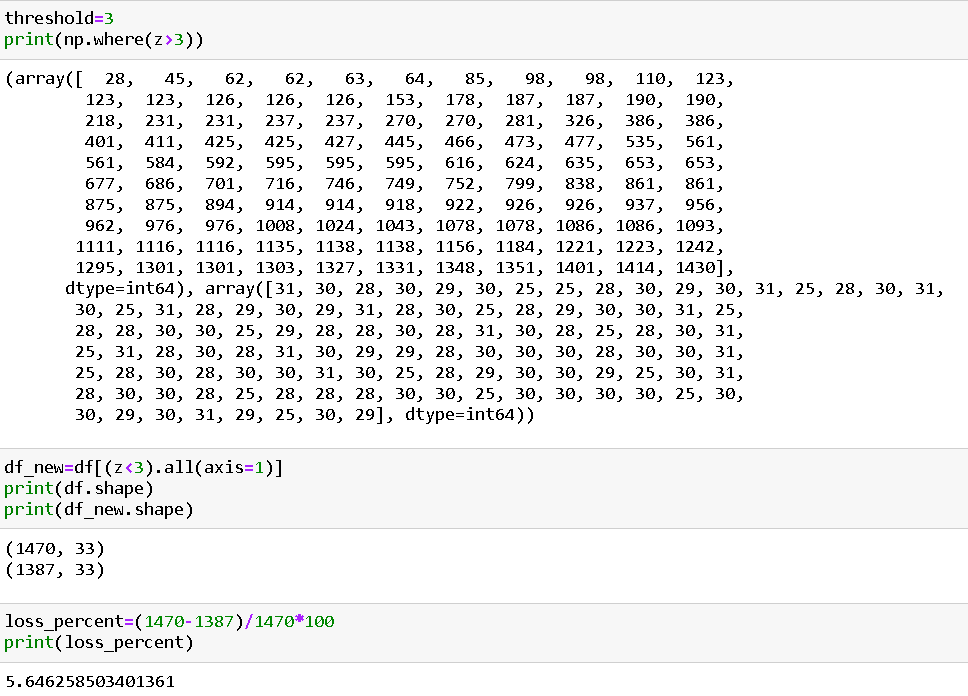
**Checking outliers:**

****

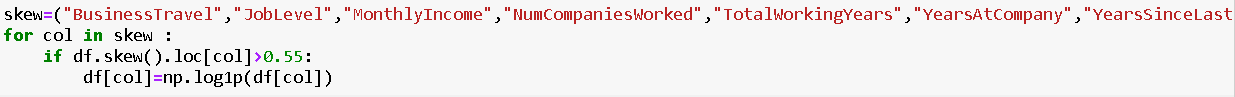
**The columns**

Monthly Income, Num companies worked, Stock option level, Total working year, Training times last year, Years at company, years in current Role, years since last promotion, years with current manager and education field has outliers. Let’s remove it to increase the model performance

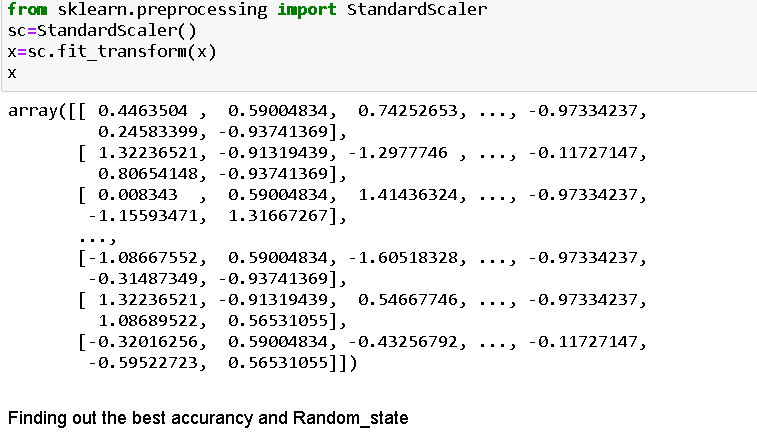
****

****

**Removing Skewness**

****

Now We will work on Scaling, training, testing, validating and hyper tunning the Model.

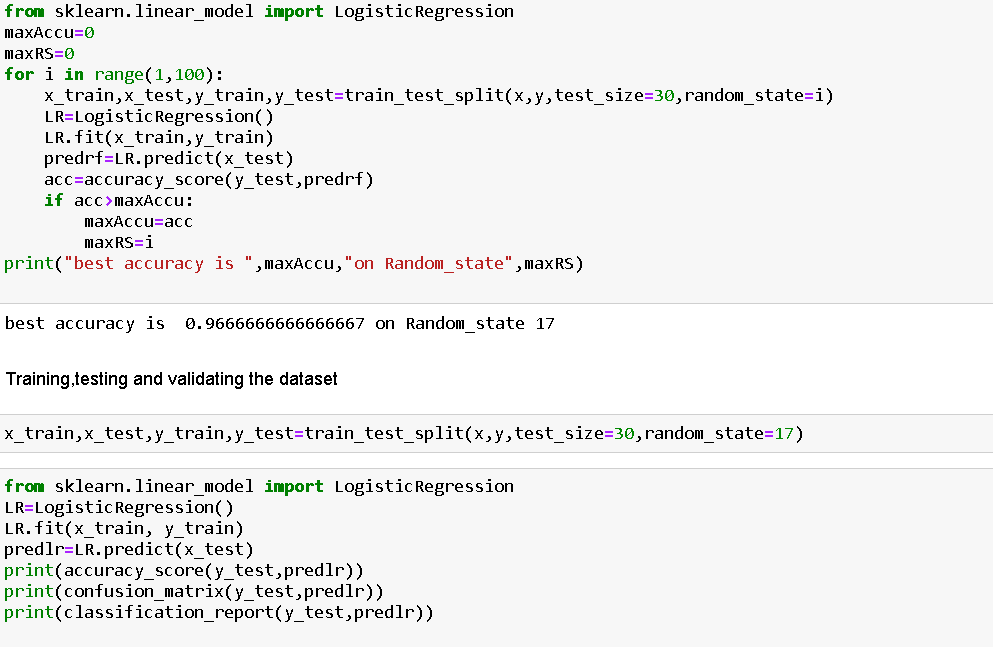
****

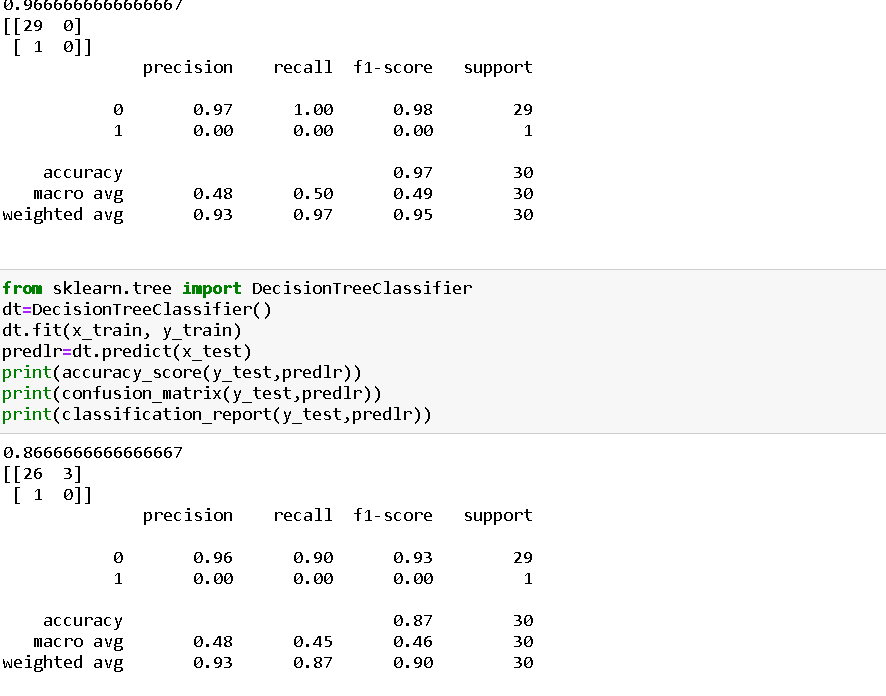
**Model Building**

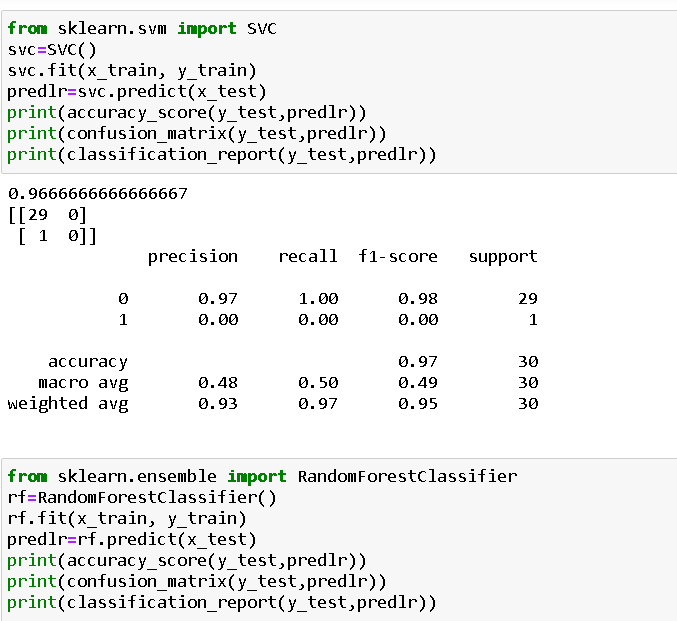
The modelling process consists in selecting models that are based on various machine learning techniques used in the experimentation. In this case various predictive models were used such as those based on decision tree, Random forest, logistic regression and SVM. The goal is to identify the best classifier for the analysed problem. Each classifier must therefore be trained on the featured set and the classifier with the best classification results is used for prediction. The classification algorithms taken into consideration are: • Logistic Regression classifier, • Decision tree classifier, • Random forest classifier, • Support Vector Machines (SVM) classification. After identifying the objectives and adequately preparing and analysing the dataset to be used, we proceeded with the design of the prediction model to identify employees that would potentially leave the company. In the construction phase of a model that implements a supervised learning algorithm, it was necessary to have a training-set available that consisted of instances of an already classified population (target), in order to train the model to classify new observations, which will constitute the test-set (in which the attribute representing the class was missing). Then, the model must be trained on a consistent number of observations in order to refine its prediction ability. The precision of the machine learning algorithms increases with the amount of data available during training.

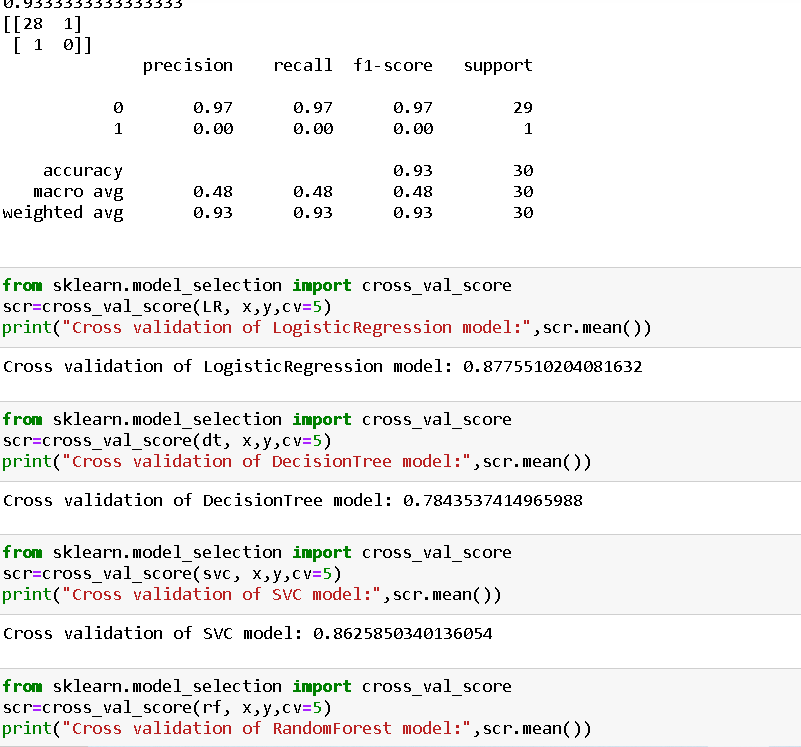
To begin, let’s split the dataset into training and test sets using 70/30 split; 70% of data will be used to train the model and the rest 30% to test the accuracy of the model. Then we can up sample the minority class, in this case the positive class.

In this step, we will start modifying model parameters, perform feature engineering and balancing data strategies to improve the performance of the models. Try with more trees in the Random Forest model, include new variables, penalize wrong predictions from the minority class until you beat the performance of our current best model.

****

****

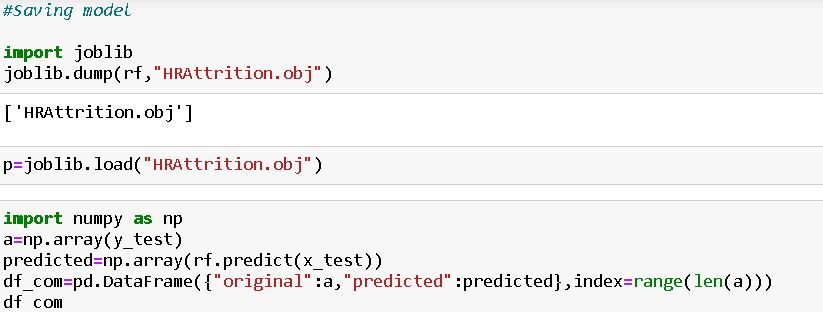
****

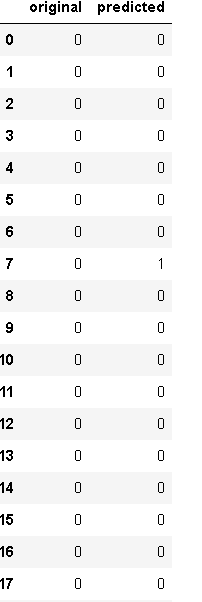
****

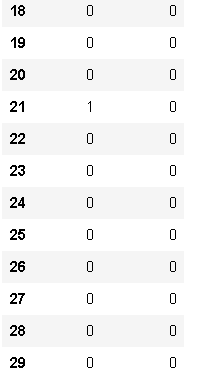
**Model validation**

Finally, after testing our models with the test set, we concluded that best model was the Random Forest (RF). Now we will Hyper tune our model with the help of GridSearchCV to increase our model accuracy.

****

****

****

****

**Wrapping up**

In this blog article we have detailed the various steps when implementing analytics use case in HR, employee attrition. We used the python to prepare the data, train different models, compare them and chose the best. With the model predictions, we created a model that would help any HR manager to retain the best talent by applying the correct strategies. This step-by-step blog article is just an example of what analytics can do for your business, and of how easy is to do it with the proper tool.