**Understanding Human Resource Attrition**

**through HR Analytics and Machine Learning**

**Sahil Kumar | April 3, 2022**

1. **Introduction**

According to IBM 2020, “Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy”.

In the field of data science, it has proved to be an important component in designing algorithms for predictions or classifications using statistical methods, and using the results to uncover key insights, which are in turn help in decision making within applications and businesses.

The basic concept of machine learning in data science involves using statistical learning and optimization methods that let computers analyze datasets and identify patterns. Machine learning techniques leverage data mining to identify historic trends to inform future models. (UC Berkeley 2020)

The typical supervised machine learning algorithm consists of (roughly) three components:

* **A decision process:** A recipe of calculations or other steps that takes in the data and returns a “guess” at the kind of pattern in the data your algorithm is looking to find.
* **An error function**: A method of measuring how good the guess was by comparing it to known examples (when they are available). Did the decision process get it right? If not, how do you quantify “how bad” the miss was?
* **An updating or optimization process**: Where the algorithm looks at the miss and then updates how the decision process comes to the final decision so that the next time the miss won’t be as great. (UC Berkeley 2020)

According to NVIDIA 2018, the different machine learning models include the following:

* **Supervised learning**, where the dataset being used is pre-labeled and classified by users to allow the algorithm to see how accurate its performance is
* **Unsupervised learning**, wherein the raw dataset is unlabeled, and an algorithm identifies patterns and relationships within the data without help from users
* **Semi supervised learning** is where the dataset contains structured and unstructured data which guides the algorithm on its way to making independent conclusions. The combination of the two types allows the algorithms to learn to label unlabeled data
* **Reinforcement learning** involves the dataset using a rewards/punishments system to offer feedback to the algorithm, thereby allowing it to learn from its own experiences by trial and error.

Building and using machine learning models involves data analytics, which “is the discovery and communication of meaningful patterns in data” (GeeksforGeeks, 2021). It involves strong reliance on the application of statistics, computer programming, and operation research to qualify performance, and favors data visualization to draw and communicate insights from the data, with the goal of smarter, better decision making and business outcomes.

**Data analytics** is usually classified into **predictive analytics** (primarily involving the use of data to make predictions about a future event), **descriptive analytics** (which looks at data and analyzes past events for insights on how to approach future events), **prescriptive analytics** (which automatically synthesizes big data, mathematical science, business rules, and machine learning to make a prediction, and then suggests a decision option to take advantage of the prediction), and **diagnostic analytics** (allows the use of historical data over other data to answer any question or provide solutions of a problem).

According to Gartner, HR analytics, also known as people analytics, is the collection and application of talent data to improve critical talent and business outcomes. HR analytics leaders enable HR leaders to develop data-driven insights to inform talent decisions, improve workforce processes and promote positive employee experience. It uses behavioral data to understand how people work and change how companies are managed. (MIT 2018). It has become a strategic tool in analyzing and forecasting Human related trends in the changing labor markets, using Career Analytics tools. (Sela, A., Chalutz Ben-Gal, Hila 2018)

1. **Problem Definition**

AIHR (Academy to Innovate HR) describes **employee attrition** as when an employee leaves the company through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or even illness or death. Whenever anyone ceases working for the company for any reason and is not replaced for a long time (if ever), that would be employee attrition. Now, **attrition** may be of two types, **voluntary** (when an employee chooses to leave the company) and **involuntary** (when the company decides to part ways with an employee).

Every year a lot of companies hire several employees. The companies invest **time and money in training** those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the **effectiveness of their employees**. **HR analytics** helps to **improve employee performance** and therefore getting a **better return on investment**. It aims to provide insight into each process by **gathering data** and then using it to **make relevant decisions** about how to **improve these processes**.



In general, relatively **high attrition** is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees. 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.

There are several ways, according to AIHR, to prevent voluntary attrition, including the following:

1. Get your managers the training they need to manage employees effectively
2. Conduct a salary survey and benchmark all your salaries
3. Conduct stay interviews
4. Revise your benefits and perks to offer ones that your employees like
5. Consider allowing more flexibility
6. Take care to hire the right people in the first place
7. Have accurate job postings
8. Remember to promote from within

According to AIM 2016, by understanding HR attrition through analytics and predictive models, companies can make the following helpful decisions:

Through predictive algorithms, companies gain better understanding and can undertake preventive measures for employee attrition. The accuracy of the model is directly proportional to the selection of parameters, which in turn, leads to the generation of the ‘type’ of predictive model most suitable for the organisation.

The dataset used for in this article for the purpose of understanding the attrition and building a predictive model on whether an employee will leave the company has “**Attrition**” as the target classification label, with the following features (or the parameters used in building the predictive model):

|  |  |  |  |
| --- | --- | --- | --- |
| **Features / Parameters** | | | |
| Age | Business Travel | Daily Rate | Department |
| Distance From Home | Education | Education Field | Employee Count |
| Employee Number | Environment Satisfaction | Gender | Hourly Rate |
| Job Involvement | Job Level | Job Role | Job Satisfaction |
| Marital Status | Monthly Income | Monthly Rate | Number of Companies Worked |
| Over18 | Over Time | Percent Salary Hike | Performance Rating |
| Relationship Satisfaction | Standard Hours | Stock Option Level | Total Working Years |
| Training Times Last Year | Work Life Balance | Years At Company | Years In Current Role |
| Years Since Last Promotion | Years With Current Manager | - | - |

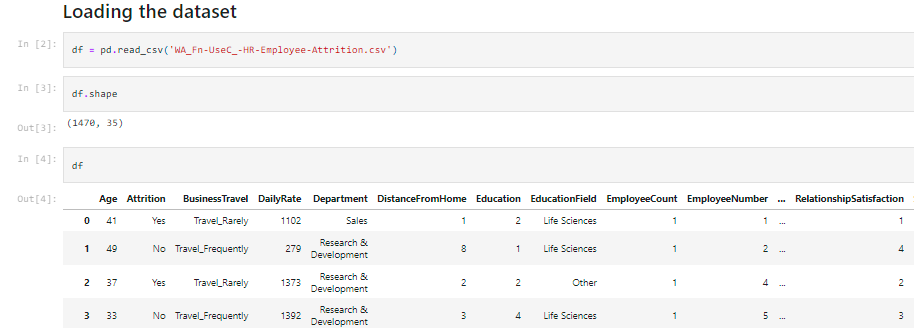
The aim of the model built and used is to predict whether an employee will undergo attrition or not. Attempts have bene made to ensure highest accuracy score of the prediction model, with high ROC AUC scores and precision scores.

After importing the necessary libraries to initiate the model building, and loading the dataset, Exploratory Data Analysis (EDA) was undertaken, followed by performing a check of data imbalance on the target column, i.e., attrition. Data imbalance was clear from the check performed, which was later fixed using SMOTE. Columns with records not relevant for building the classification model were dropped, which was followed by encoding of all categorical data. Relationships between the various features, and with the target column, were visualized using data visualization techniques, followed by establishing correlations, checking for, and removal of outliers and skewness, and using SMOTE to balance the dataset. The model building involved checking the accuracy and performance of prediction using logistic regression, random forest classifier, decision tree classifier and support vector classifier.

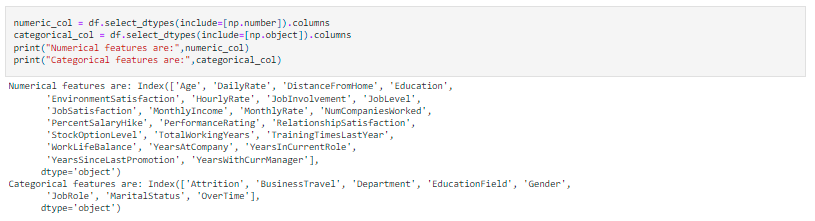
1. **Data Analysis**

IBM (2020) explains Exploratory Data Analysis (EDA) as a process used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. The dataset for the model in this article was downloaded in CSV format from:

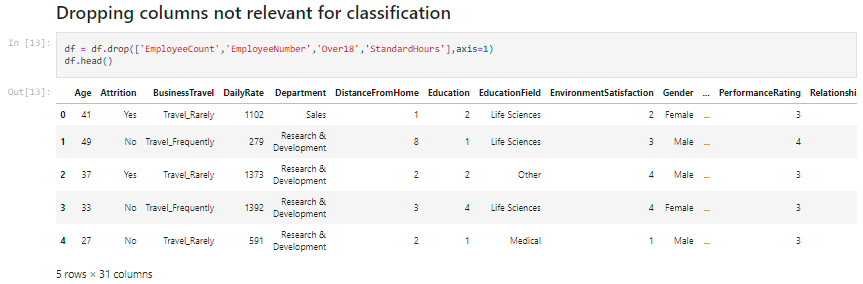
<https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>



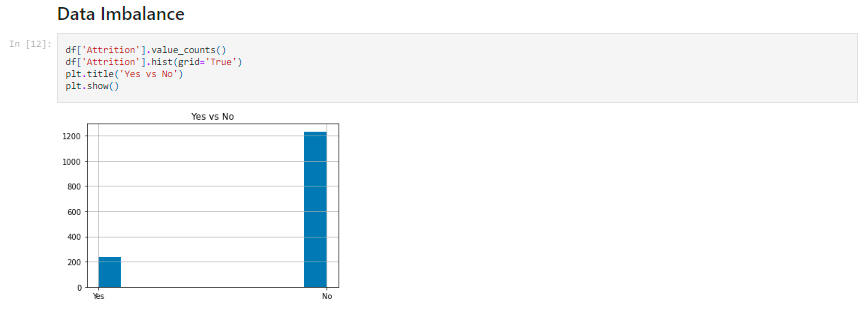
The records in the dataset were a mix of both categorical (9 columns with object data type) and numerical data (26 columns with int64 data type), having 35 columns and 1,470 records (rows).



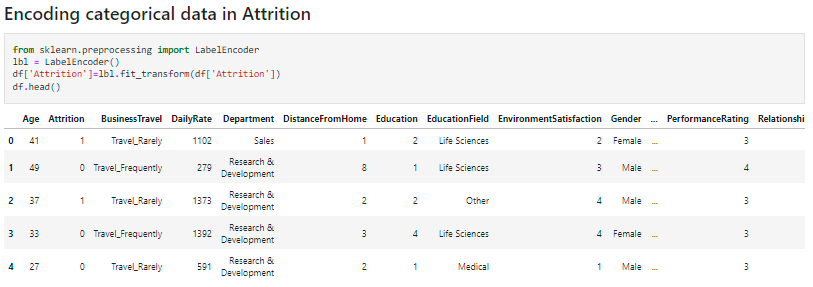
The dataset was checked for null values, and none were identified. The number of unique records in each column were identified using **df.nunique()** , which revealed 4 columns irrelevant for the prediction namely, EmployeeCount with 1 unique value, EmployeeNumber with all 1470 unique values, Over18 with 1 unique value, and StandardHours, again with 1 unique value. These columns were dropped.



On checking for data imbalance on the target column (attrition) using **value\_counts()** and plotting a graph for the same, it was noted that the dataset was imbalanced, with records demarcated “Yes” for attrition being very low as compared to those demarcated “No” for attrition, implying low attrition numbers in the dataset. There were 1233 (83.88%) cases where employees did not undergo attrition and 237 (16.12%) cases where employees underwent attrition.



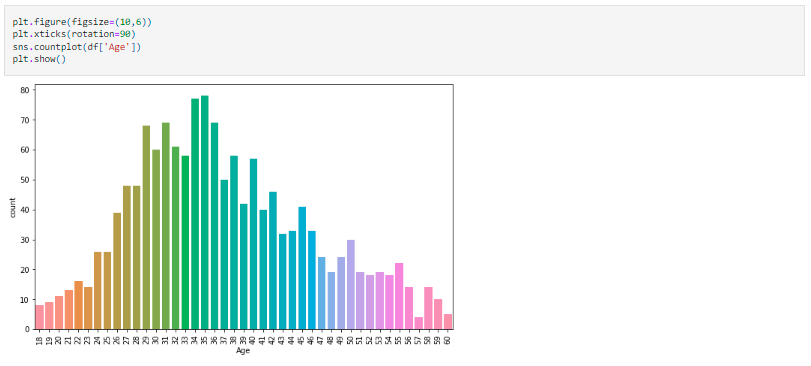
The data in the target column – attrition was encoded using the Label Encoder, making 1 = Yes and 0 = No.



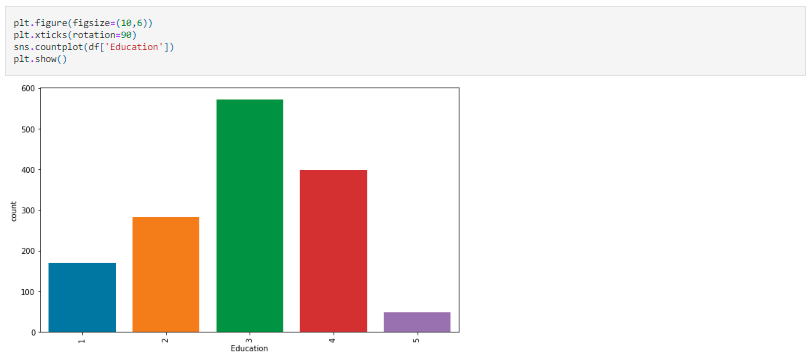
Features and target were then visualized using various data visualizations to understand the relationships between the target and features, as well as among the features.

**Visualizing the relationships**

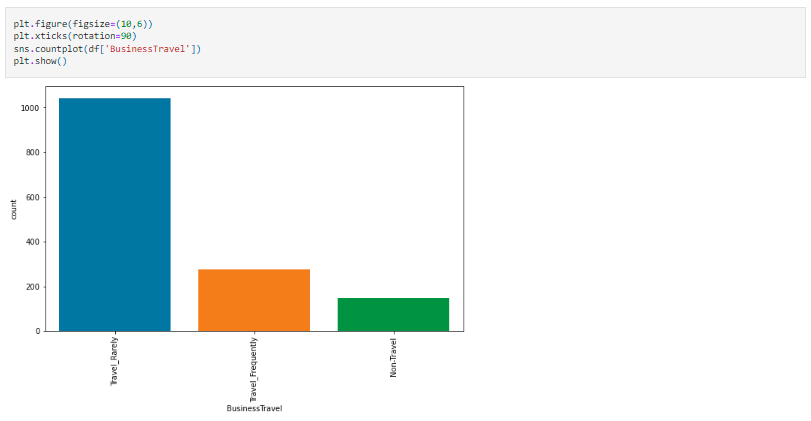
1. **Distribution of age** – most employees falling between the ages of 26 and 45



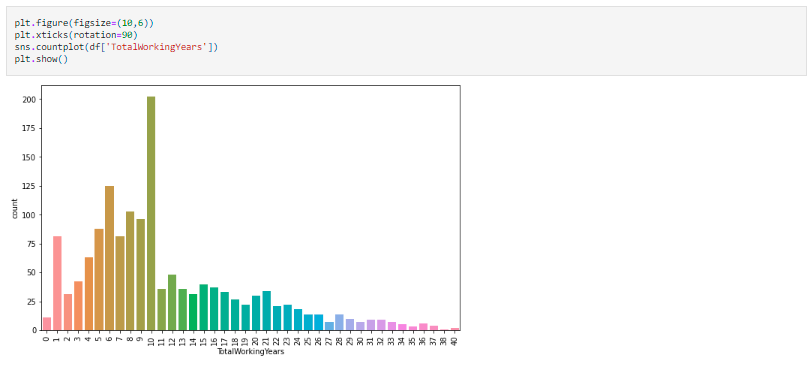
1. **Level of education** – most employees have completed level 3 of education in the company



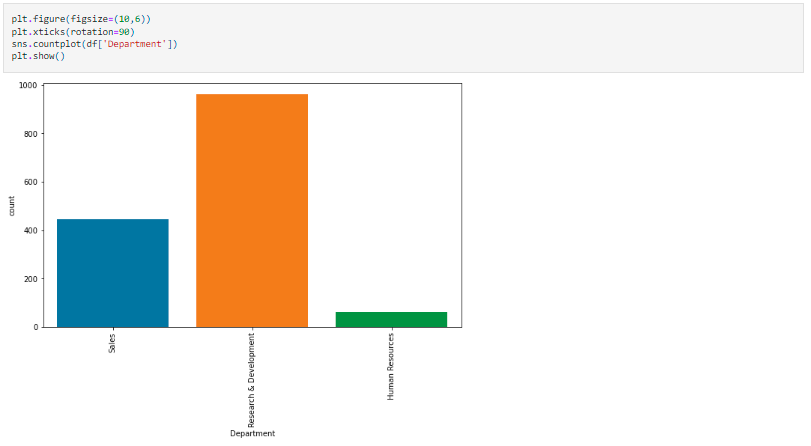
1. **Frequency of travel** – most employees travel rarely; more than 200 travel frequently; and less than 300 do not travel or are not required to travel as a part of their job profile / description.



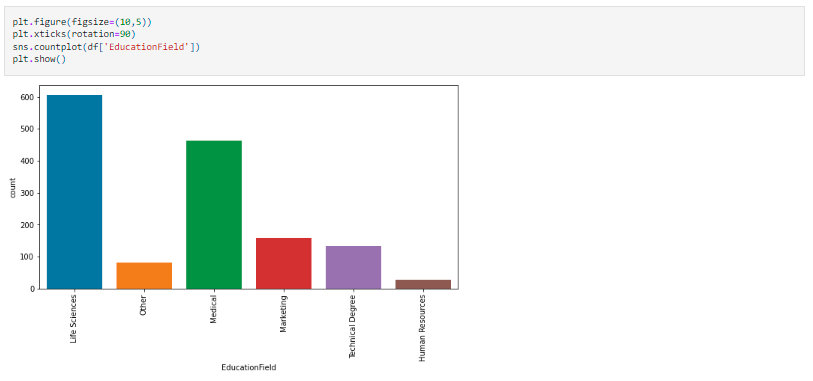
1. **Total number of working years** – most number of employees (around 200) have 10 years of total work experience, followed by those who have 6 years of total work experience (around 125), followed by the rest.



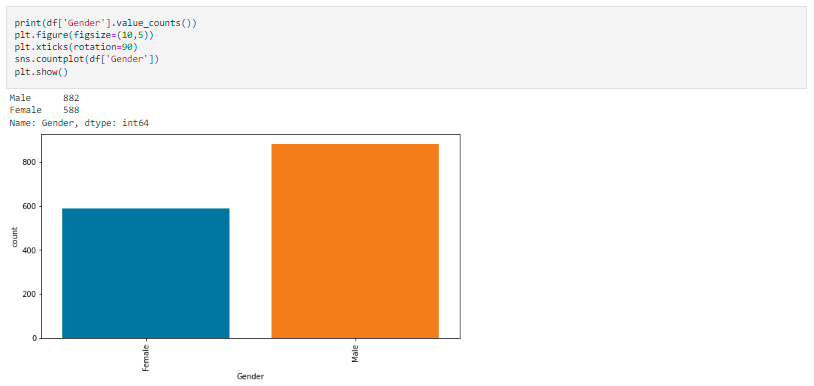
1. **Department** – most number of employees work with the Research and Development team, followed by those in Sales, and finally in the Human Resources teams.



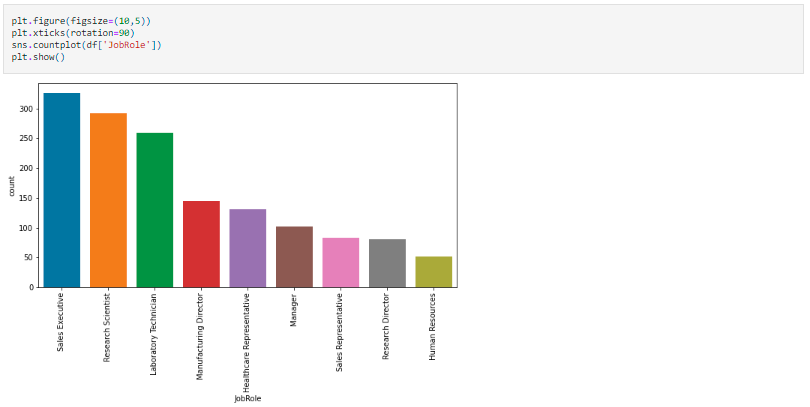
1. **Field of education** – Most employees' education field is in the life sciences, followed medical, marketing, technical. Least have a background in HR. The remaining are classified as Others.



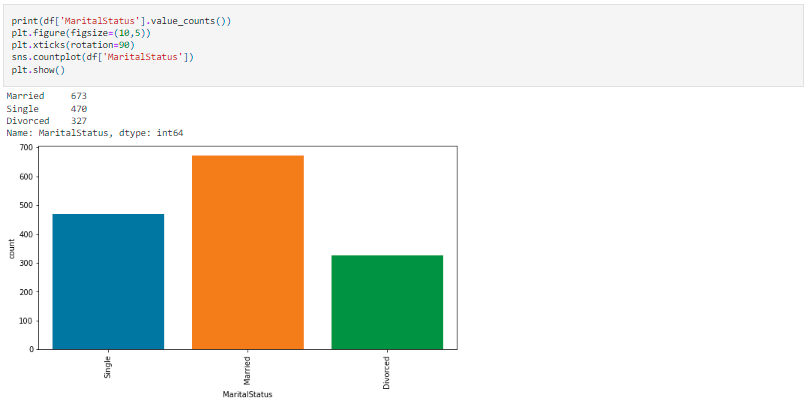
1. **Gender** – There appears to be an imbalance in terms of gender at the organization as there are 882 males and 588 females only.



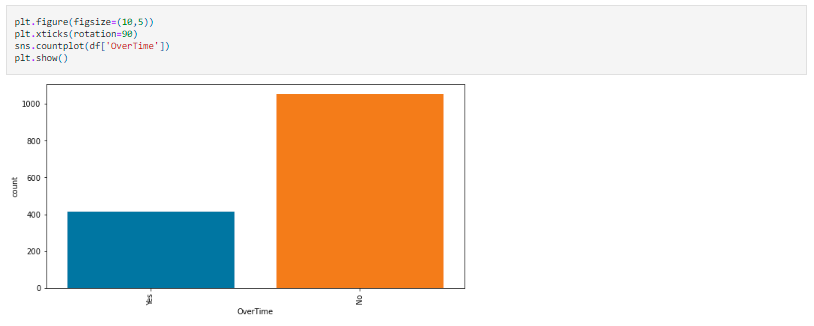
1. **Job role** – Most number of employees are employed as Sales executives, followed by research scientists, Lab technicians, Manufacturing director, Healthcare reps, manager, sales reps, research director and human resource execs.



1. **Marital status of employees** – Most employees are married, followed by single status, and finally divorced

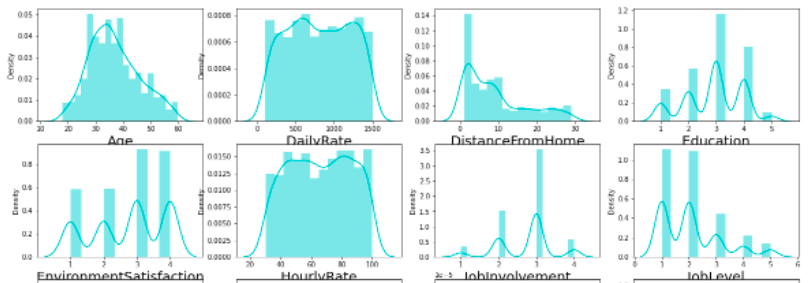


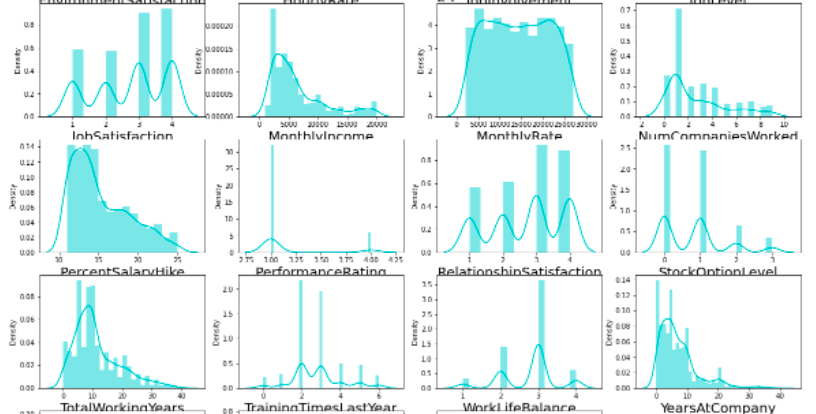
1. **Overtime** – It is evident that most employees do not or have not worked on overtime.

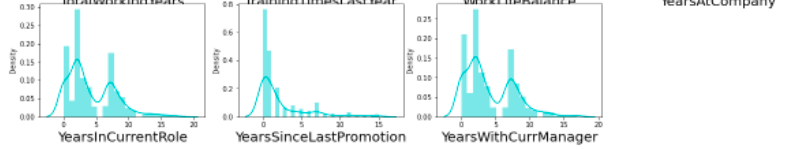


1. **Distplot to check for distribution of data** – Most columns in the dataset are not normally distributed and have lot of skewness

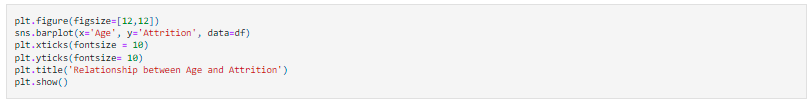


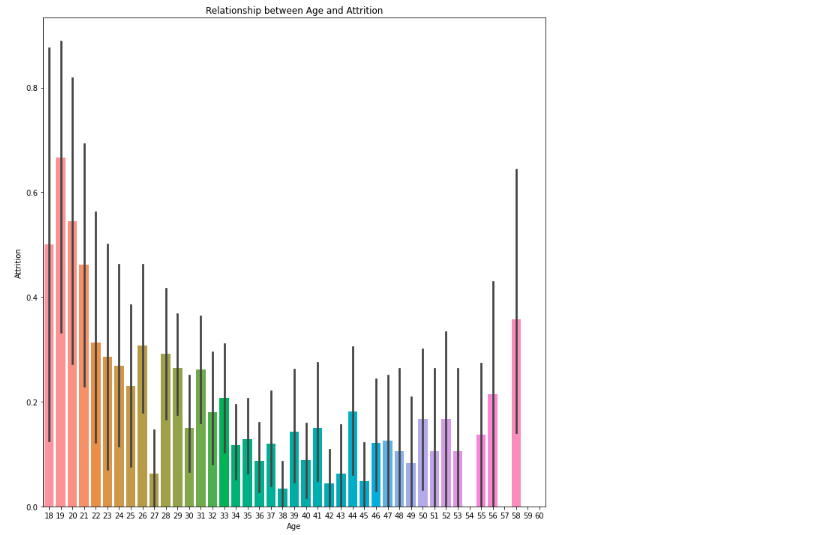




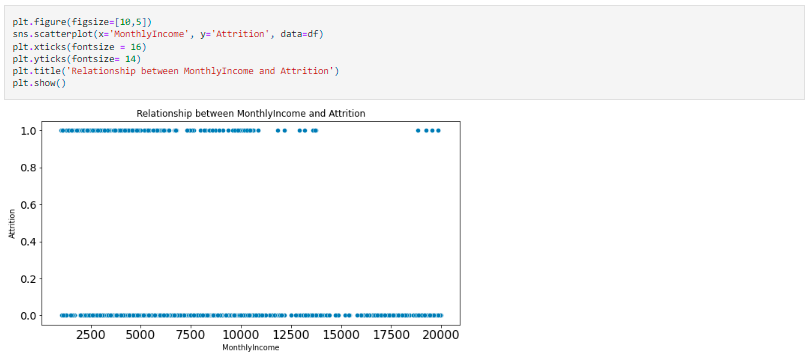


1. **Relationship between Age and Attrition** – The highest attrition rate is among employees in the ages of 18 - 21, which keeps declining inconsistently, but eventually sees a rising trend as on approaches the ages of 50 and above.

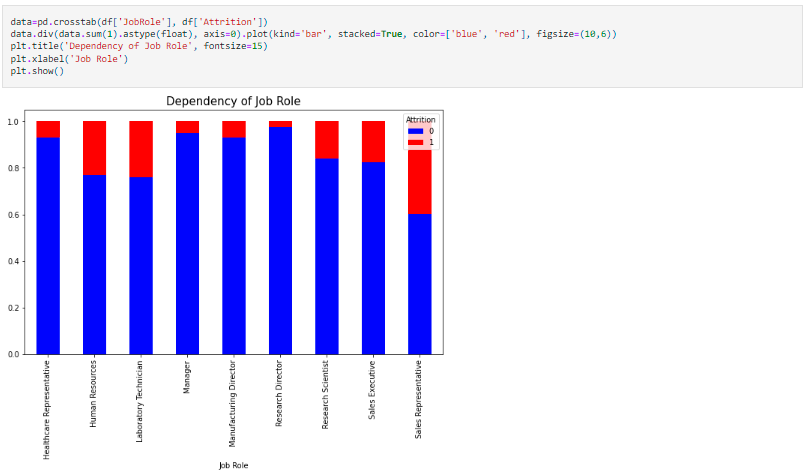




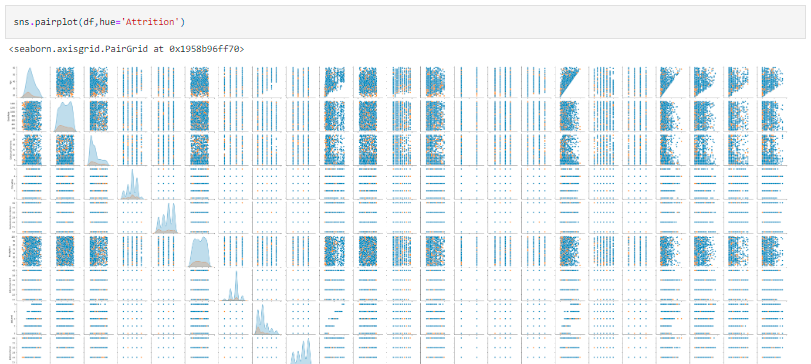
1. **Relationship between Monthly Income and Attrition** – Employees with lower monthly income quit more



1. **Relationship between Job Role and Attrition** – There is very high attrition among Sales reps, followed by HR and Lab Techs, and finally Research scientists and sales execs.



1. **Pairplot with hue as “Attrition”** – It visualizes the relationship of each feature with another. The plots on the diagonal reflect the distribution of data (partial screenshot below)



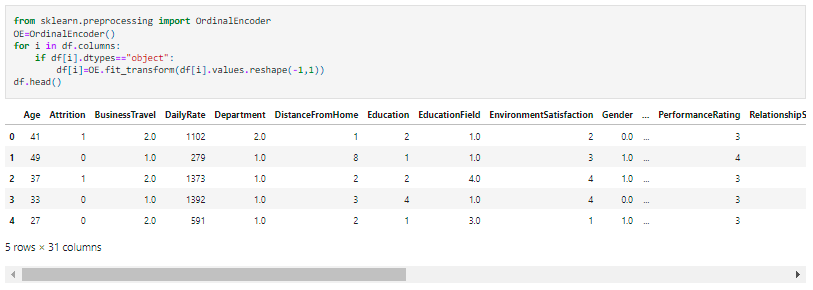
1. **EDA Concluding Remarks**

After completing the Exploratory Data Analysis (EDA), we identified the relationships between the various features and the target, as well as the various distribution of records across the various features. The dependent variable, or the target, i.e., attrition was imbalanced. The sample size for building an effective prediction model for classification of whether an employee will undergo attrition, i.e., leave the company, was too small. The dataset contains information from only few departments from the organization. There were no null values present, however, based on the number of unique values present in each feature, some columns were identified as irrelevant and dropped from the dataset for building an effective prediction model.

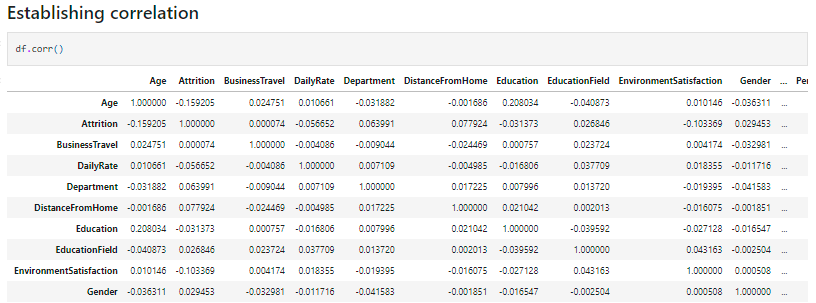
1. **Pre-Processing Pipeline**

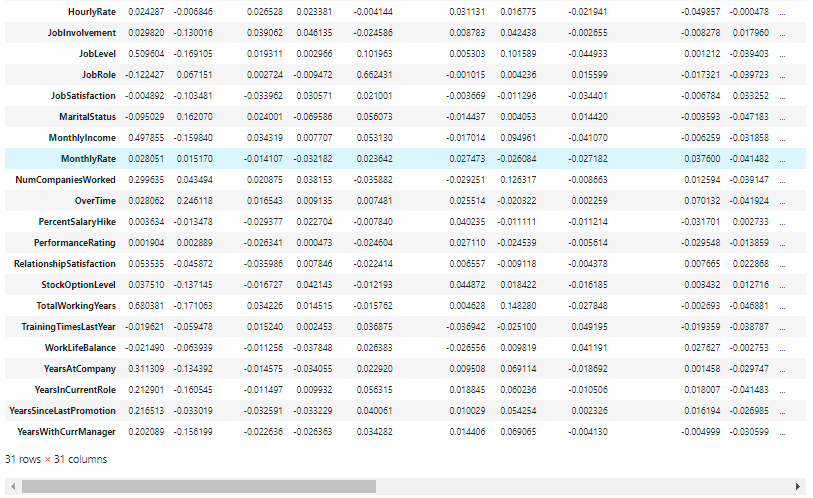
According to Medium (2020), “Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. Incomplete, noisy, and inconsistent data are the inherent nature of real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.”

Upon completion of sufficient exploratory data analysis and visualization of relationships between features and the target, all the categorical data in the dataset was encoded using an Ordinal Encoder, to make the records ready for establishing correlations between features and target, checking for outliers and their subsequent removal, checking for skewness, scaling the data, fixing the data imbalance using oversampling (SMOTE), and finally, moving on to the building the predictive model.



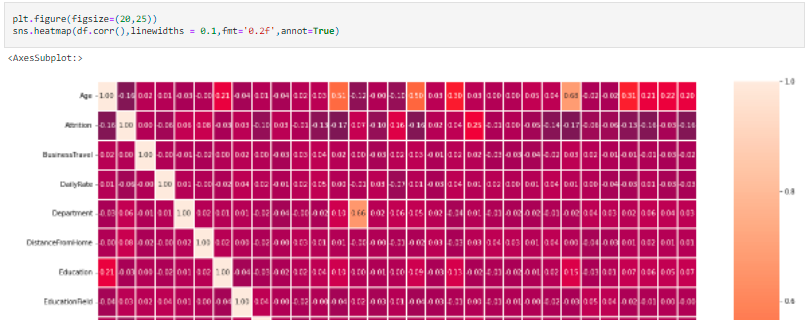
After encoding the categorical data, we established the correlations among the features, and between features and the target (attrition).





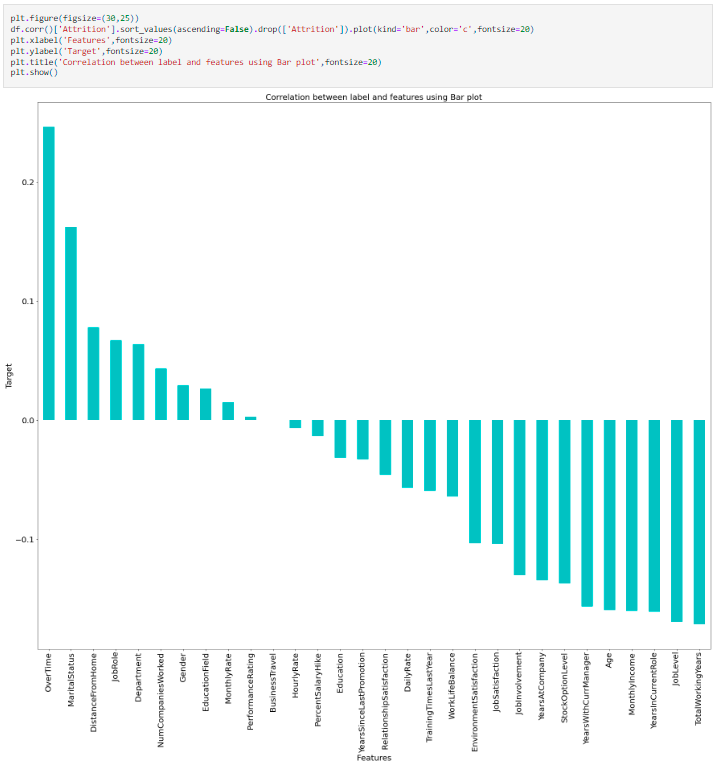
We identified the positive and negative correlations, along with their magnitude (highly positive, highly negative, etc.). We observed that there is very high positive correlation with overtime and negative correlation with total working years.

Moreover, a heatmap was plotted to clearly establish and visualize the correlations (partial snapshot below).



Further, a bar plot was visualized to establish the correlation between the features and the target, as below, and the following insights were obtained:

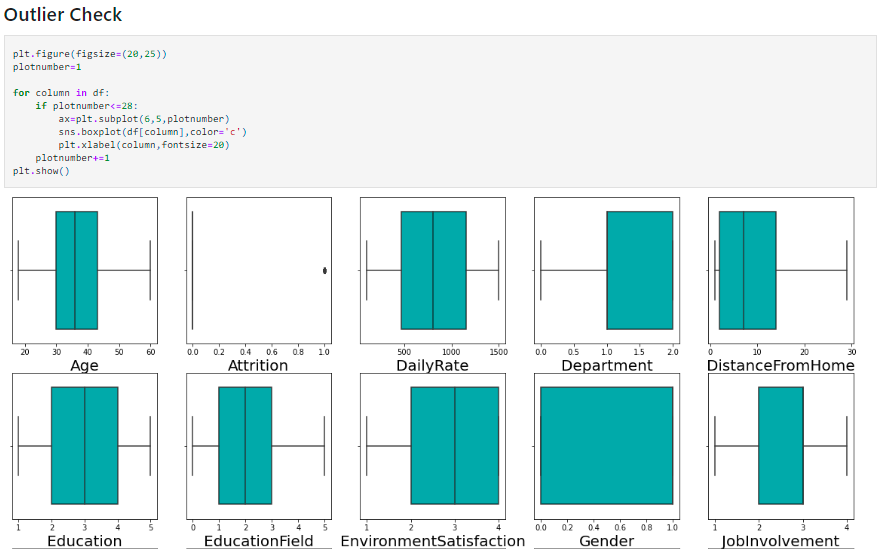
* There existed a highly positive correlation between Attrition and overtime, and attrition and marital status
* A highly negative correlation of attrition was observed with total working years, job level, years in current role, monthly income, age, years with current manager, stock option level, years at company, job involvement
* Least amount of correlation with attrition was observed with performance rating, business travel and hourly rates

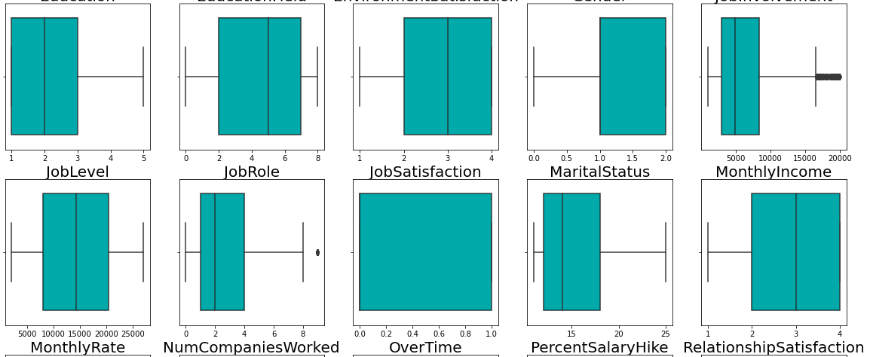


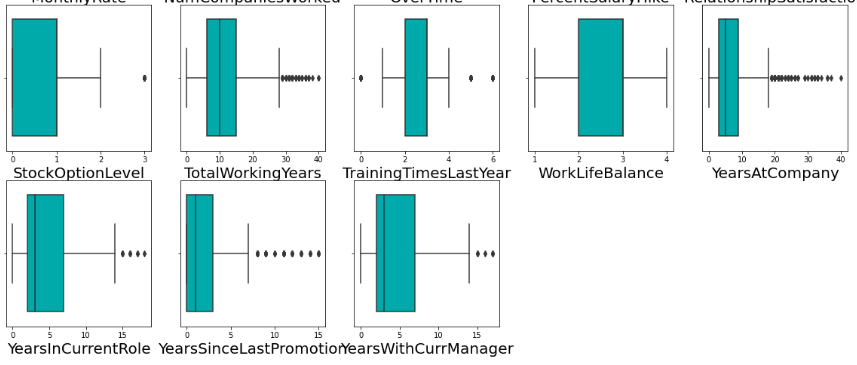
Based on the findings above, the columns with negligible correlation were dropped.



This was followed by a check for outliers, and their subsequent removal using z-score.

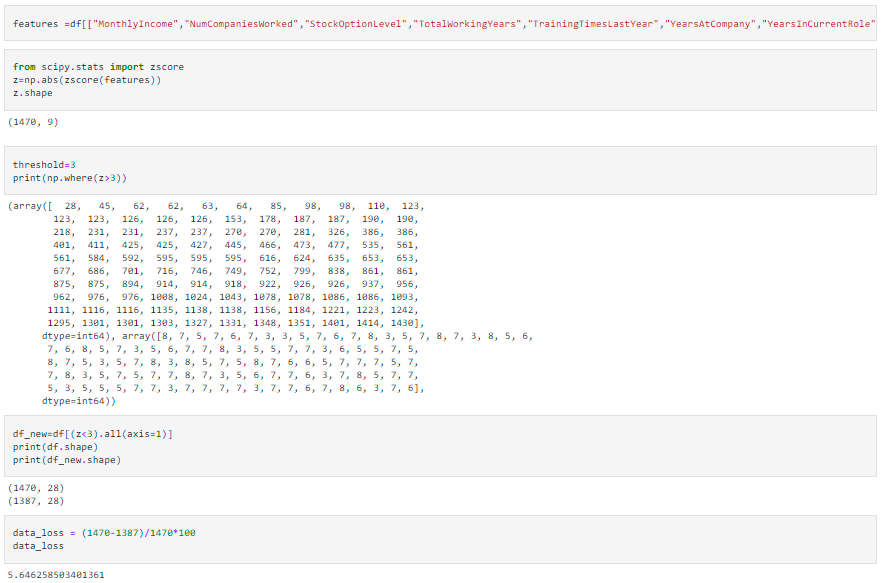




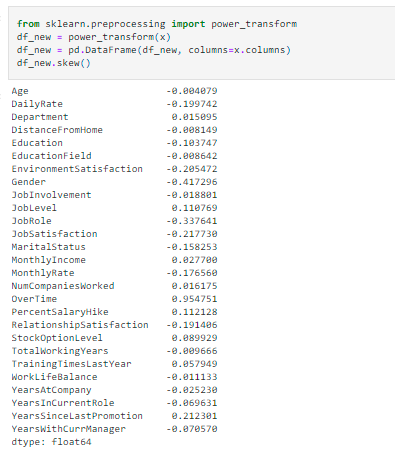


Outliers were identified in a few columns including “MonthlyIncome”, "NumCompaniesWorked", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion" and "YearsWithCurrManager".

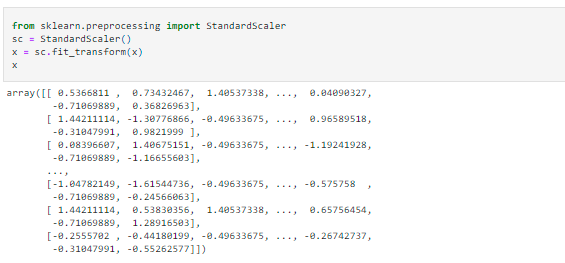
Outliers were removed using z-score since the data loss was within acceptable limits as follows:



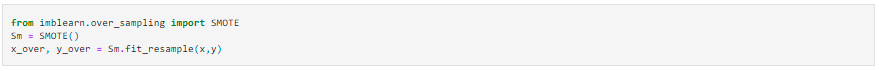
The dataset was split between features and target and checked for skewness. The observed skewness was removed using Power Transformer.

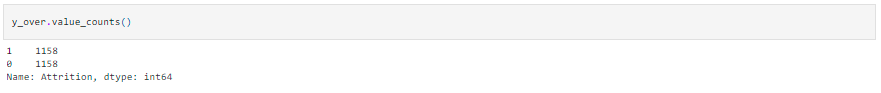
 

After removing the skewness, Standard Scaler was used to achieve feature scaling.



The final step before building the prediction model was overcoming class imbalance using SMOTE (Synthetic Minority Oversampling Technique), which is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together. (Satpathy, Analytics Vidhya 2020).

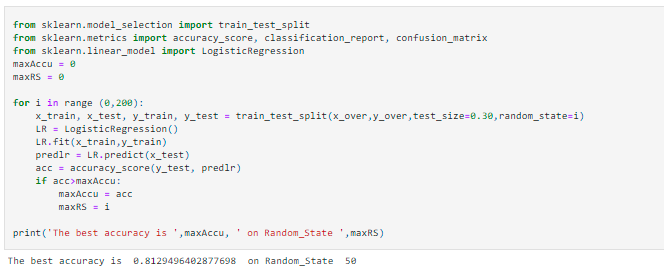




1. **Building Machine Learning Models**

The model building for predicting attrition of an employee involved using classification machine learning algorithms, including Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Classifier:

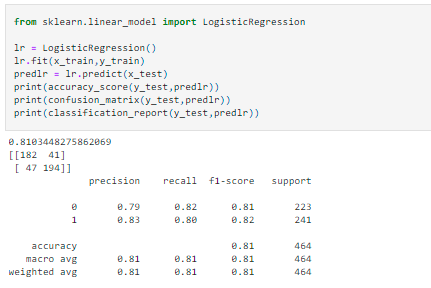
1. Best Accuracy on Max Random State



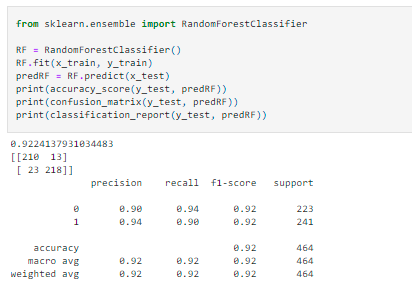
1. Train Test Split



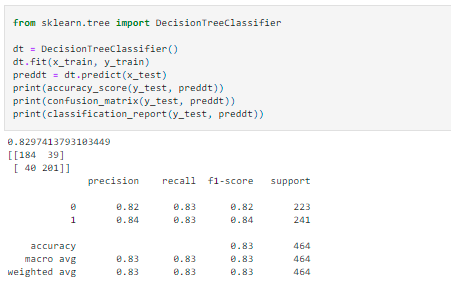
1. Model testing: Logistic Regression



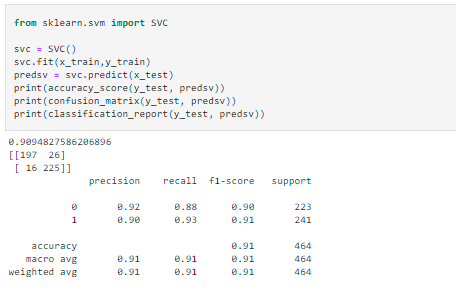
1. Model testing: Random Forest Classifier



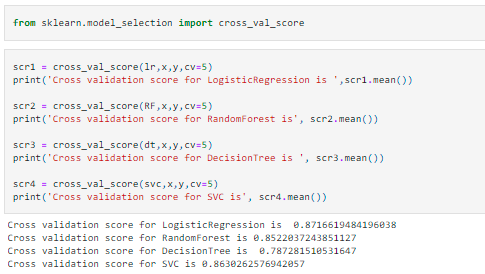
1. Model testing: Decision Tree Classifier



1. Model testing: Support Vector Classifier (SVC)

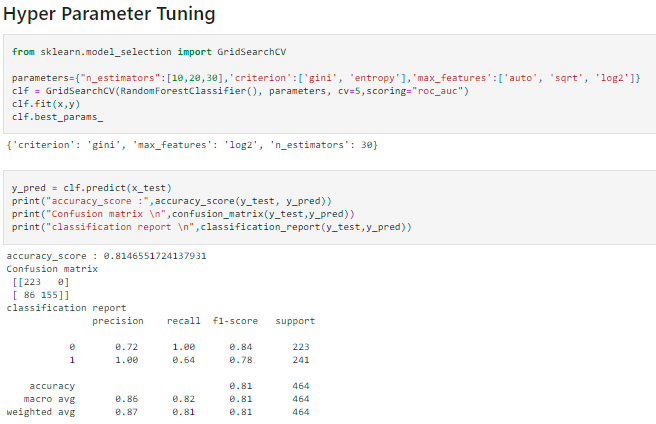


1. Model Cross Validation



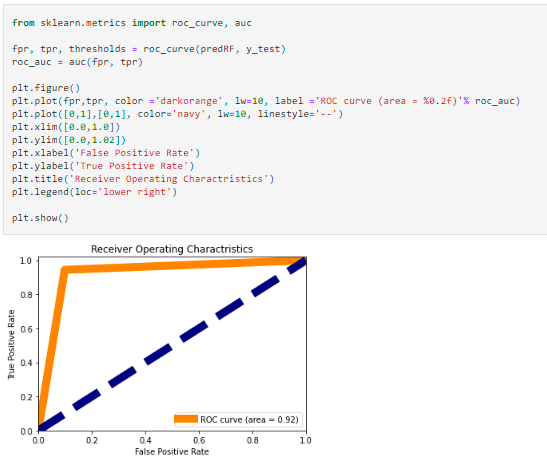
Based on the accuracy scores and model cross validation, Random Forest was the best model.

1. Hyperparameter Tuning



Post hyperparameter tuning, the Random Forest Classifier presented an 81% accuracy of predicting the attrition of employee with certainty.

1. Visualizing the ROC (Receiver Operating Characteristic) curve



The area under the ROC curve presented at 0.92, which is a decent and high result.

1. Saving the prediction model



1. **Concluding Remarks**

The project has built a machine learning algorithm using the Random Forest Classifier, to predict whether an employee will leave the organization (attrition) based on pre-defined parameters (features) available in the dataset. Four different classifiers were used to test their accuracy and effectiveness, and the Random Forest Classifier was finalized as the best model for predicting the attrition, with an accuracy of 81% and an ROC area of 0.92. The advantage of the prediction model is that the company may be able to predict the potential attrition, thereby saving them time and cost of going through the recruitment process again and again, and may be able to use techniques, surveys, and trainings to focus on employees who may be at high risk of attrition, i.e., leaving the company. However, since the dataset was limited in size, and scope of parameters, may not be adaptable to the entire organization, or for predicting attrition in other companies, capturing different data pertaining to their employees.

1. **References**
2. <https://www.ibm.com/cloud/learn/machine-learning>
3. <https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/>
4. <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/>
5. <https://www.geeksforgeeks.org/data-analytics-and-its-type/>
6. <https://www.gartner.com/en/human-resources/glossary/hr-analytics#:~:text=HR%20analytics%20(also%20known%20as,and%20promote%20positive%20employee%20experience>.
7. <https://professional.mit.edu/course-catalog/people-analytics-transforming-management-behavioral-data>
8. <http://www.eng.tau.ac.il/~bengal/Chalutz_Career_Analytics.pdf>
9. <https://www.aihr.com/blog/employee-attrition/>
10. <https://analyticsindiamag.com/predictive-attrition-model/#:~:text=Benefits%20of%20Predictive%20Attrition%20Model&text=Minimize%20cost%20of%20new%20talent,plan%20and%20minimize%20the%20loss>
11. <https://www.ibm.com/in-en/cloud/learn/exploratory-data-analysis>
12. <https://medium.com/swlh/data-preprocessing-and-data-modeling-for-kaggle-house-price-prediction-data-in-python-c04055ded258>
13. <https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/>
14. Source of dataset: <https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>
15. Source of model built by author: <https://github.com/SahilKumar3000/DataTrainedEvaluationProjects/blob/main/Evaluation%20Project%202%20-%20HR%20Attrition.ipynb>