Human Resources Analytics - Employee Churn

1) Business Problem and Background

Hiring and Retaining new employees are extremely convoluted tasks that require capital, time and skills. Most of the times when skilled employees leave, it costs far more to replace them than providing them with incentives to retain them. As per the statistics, small businesses spend around 40 percent of their funds on processes that do not generate revenue. Also companies spend roughly 15% of an employee's salary to recruit new candidates. Therefore, our hypothetical company wants to apply data science methodologies to deal with the critical problem of Employee Churn based on factors such as job satisfaction, work life balance, performance rating, education and job involvement.

2) Data Extraction and Understanding

The data consists of the following attributes:

1. Job Involvement - measures how involved the employees are towards the company's objectives.
2. Education - the level of education of employees.
3. Education Field
4. Job Satisfaction - measures the level to which employees are satisfied with work.
5. Performance Rating - The annual rating gained by the employees.
6. Work Life Balance - measures on a scale, how well the employees are able to manage personal and professional life
7. Total Years at Company .
8. Total working hours per day.
9. Distance from Home - measures the travelling distance of employees.
10. Years in the current role.
11. Years since last Promotion/Appraisal.
12. Years with the current manager.

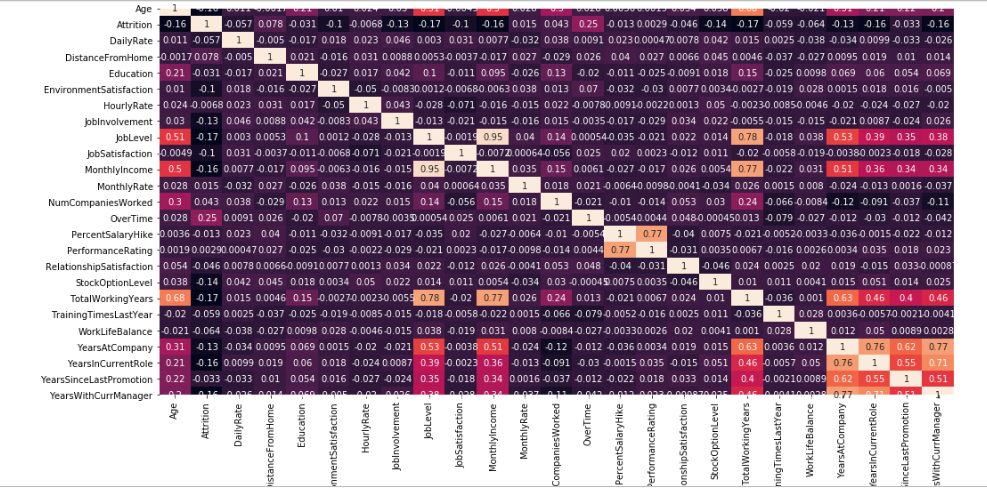
3) Exploratory Data Analysis

Exploratory data analysis is the preliminary step in data analysis. It is used to summarize the main characteristics of data, gain better understanding of data set and to unleash the various relationships in the underlying data. It can also be used to determine features that have a significant impact on the machine learning algorithm in which the predictive model will be deployed. It refers to the process of performing critical investigations on data so as to discover patterns, to search anomalies, to test hypothesis and to check assumptions with the help of statistical analysis.

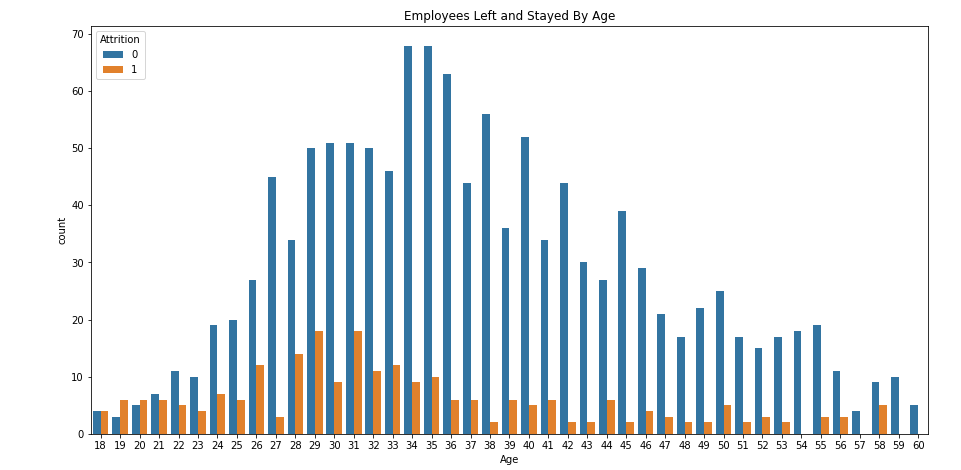
It is good to understand the data first and try to get as many insights from it as possible. Exploratory Data analysis is all about making sense in data before getting our hands dirty on it. Statistical analysis is used to carry out analysis such as mean, median and mode in the data set. The pandas data frame has various functions such as describe(), info() and head() are used to explore the initial patterns in the data.

A) First we saw the initial rows of our dataset using the pandas head() method. After that, we checked the dataset to find if it contains any missing values. In our dataset, there were no missing values and hence this handling was not needed.

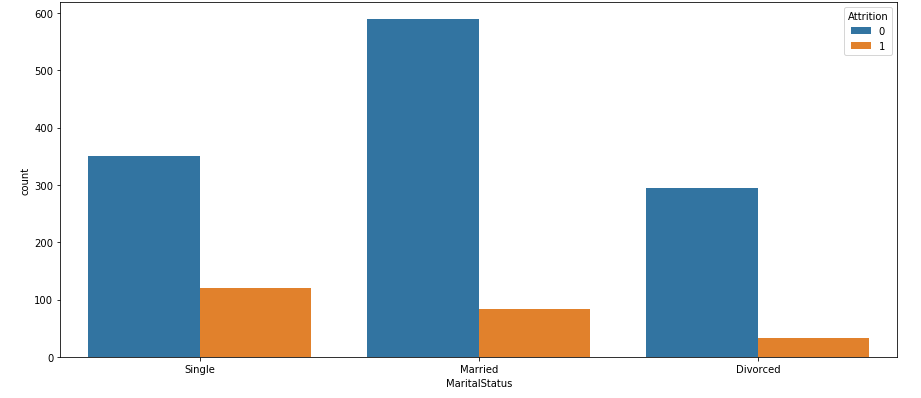
B) Next, we generate a heatmap to see the correlation among the different independent variables. After analysis, it was observed that there is a strong correlation between Total working years and age and also between Total working hours and Monthly income. Naturally, the more an employee works at a particular company, the higher the salary becomes with promotions in due course of time.

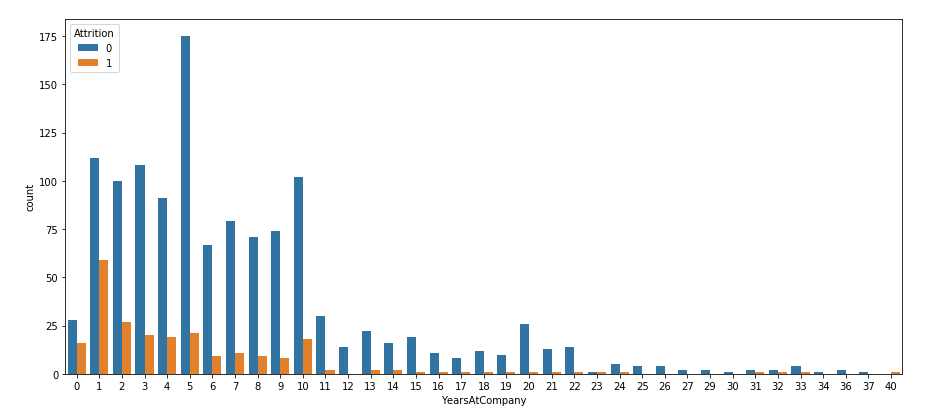


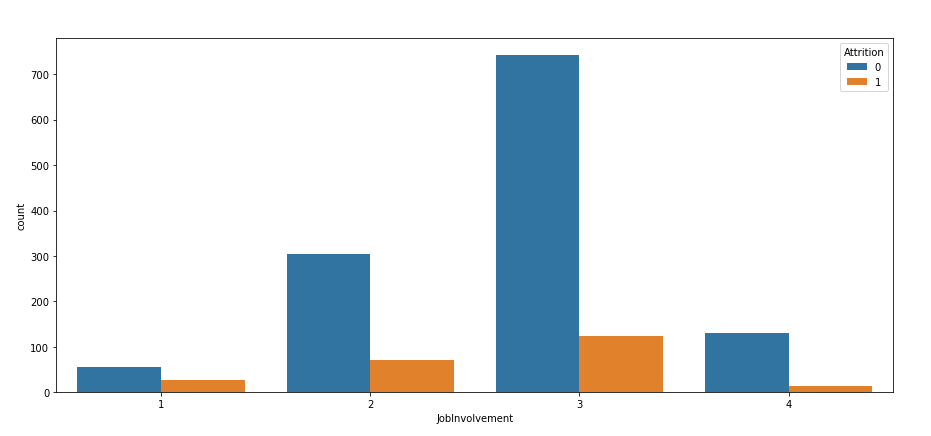
C) Using the Seaborn Countplot, let us explore the number of people who left and stayed by Age. To generate a countplot, we have to specify either one of the x or y parameters and a hue parameter that shows the count for each type of category.



Similarly, let us explore the countplot by Marital status, Job involvement and number of years at company.

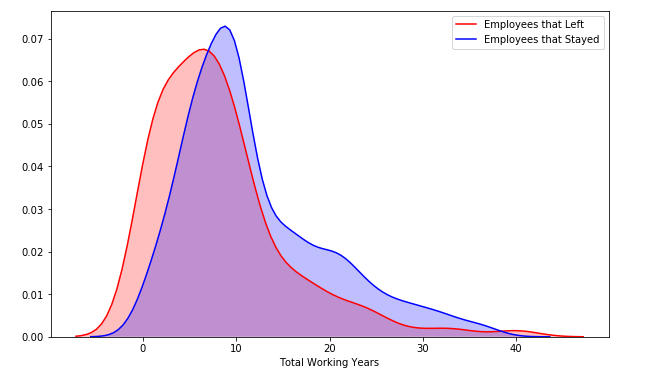
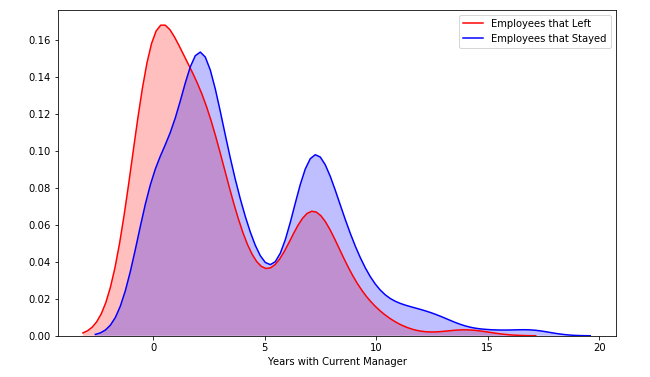
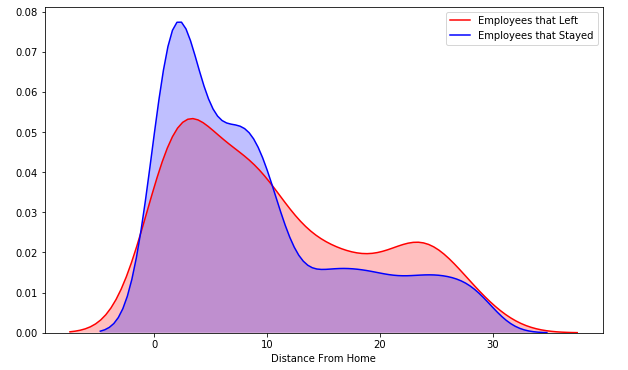






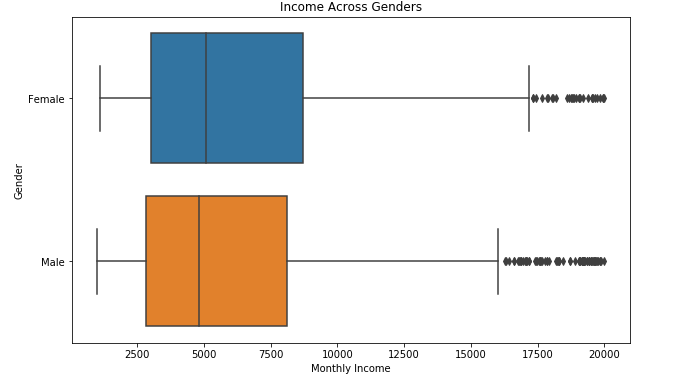
D) Now, let us explore the Kernel Density Plot for continuous variables. Kernel Density plots determine the probabilities and their distributions for continuous variables. Here we will be plotting the kdeplot using the Seaborn package for Distance from home, Years with current manager and total working hours fields.

From the first plot we can see that as the Distance from home increases, the employees are likely to quit from the organization. There can be several reasons for this behaviour.

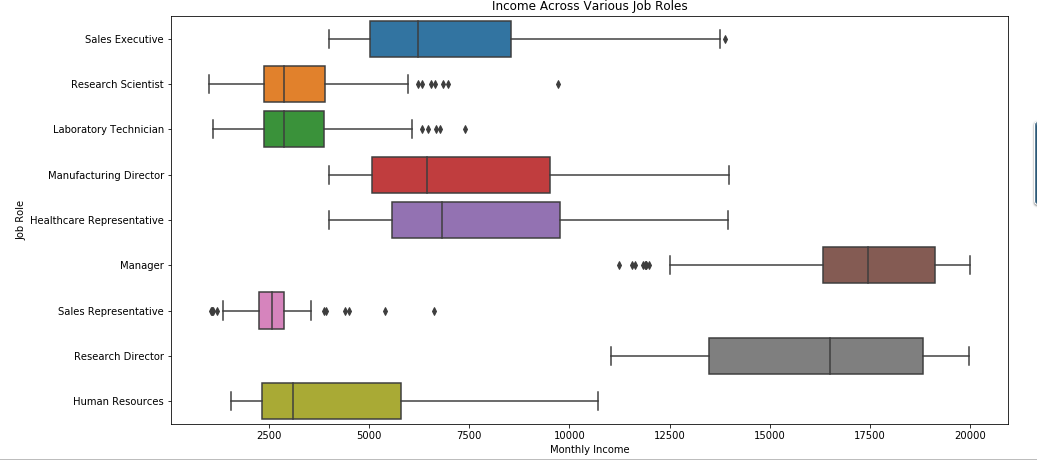


It can be seen that employees tend to quit the current job in the early years and as their period increases, the employees are less likely to leave the organization.

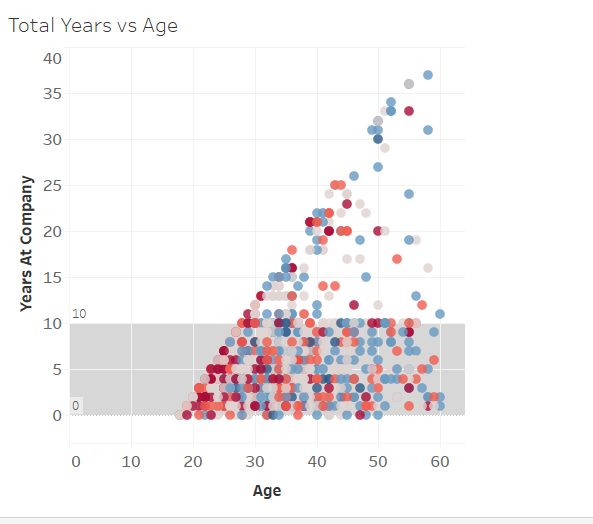
E) Now we will be plotting the box plots for Monthly income by Gender and Job Roles .



The income distribution somewhat similar for both the males and females and there is no much distinction



F) Using Tableau Public



4) Data Pre-processing.

When we think about data, we usually think about large datasets, usually comprised about a number of rows and columns. While its likely a scenario, its definitely not the case. Most of the times the data is available in unstructured formats like videos, audios, images, text, speech recordings and all.

As the data for this project was scraped from multiple websites and joined together to create a parent data frame, the quality of data should be our priority. These quality should be handled adequately before modelling the data with a machine learning algorithm.

1. Encoding Categorical Variables

There are many categorical variables in our dataset that needs to be encoded for the modeling purposes. In this project, I have used the sklearn OneHotEncoder to perform encoding of categorical variables

1. Feature Scaling/ Standardization

The different features in our dataset belong to different ranges. These differences in ranges can pose a serious threat to our analysis since there is a possibility that one feature might dominate our analysis than the rest. To overcome these problems, Standardization techniques are used to scale the features so that they all fall within a same range. As a result, all the features contribute equally to our predictive analysis. In our analysis, we used the sklearn.preprocessing.MinMaxScaler class to perform feature scaling and get the data ready for our modelling.

5) Predictive Modelling

Predictive modelling entails the use of machine learning algorithms to analyze the given dataset and anticipate the outcomes for new data. In this project, I have used the Logistic Regression, Random Forest Classification and an Artificial Neural network using deep learning to predict the employee Churn. Let us explore each of them in detail.

A) Logistic Regression Classification - Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example, To predict whether an email is spam (1) or (0)Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

We use the sklearn .linear\_model.LogisticRegression module for this purpose. We instantiate the object and then fit our model to the training set. We perform prediction using the testing set and also calculate the various classification key performance indicators such as accuracy, confusion matrix and the classification report showing the precision, recall and f-1 score. The accuracy of this module was 87 percent.

B) Random Forest Classification - Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

We use the sklearn .linear\_model. RandomForestClassifiermodule for this purpose. We instantiate the object and then fit our model to the training set. We perform prediction using the testing set and also calculate the various classification key performance indicators such as accuracy, confusion matrix and the classification report showing the precision, recall and f-1 score. The accuracy of this module was 85 percent.

C) Artificial Neural Network - Artificial neural networks are built of simple elements called neurons, which take in a real value, multiply it by a weight, and run it through a non-linear activation function. By constructing multiple layers of neurons, each of which receives part of the input variables, and then passes on its results to the next layers, the network can learn very complex functions. Theoretically, a neural network is capable of learning the shape of just any function, given enough computational power. Very effective for high dimensionality problems, able to deal with complex relations between variables, non-exhaustive category sets and complex functions relating input to output variables. Powerful tuning options to prevent over- and under-fitting. Theoretically complex, difficult to implement (although deep learning frameworks are readily available that do the work for you). Non-intuitive and requires expertise to tune. In some cases requires a large training set to be effective.

We use Tensorflow and Keras API to train our Artificial Neural Network. We add the multiple Dense layers to our Sequential model. After that we declare the cost function, optimization algorithm and then fit the neural network to our training set with a pre defined number of epochs. We instantiate the object and then fit our model to the training set. We perform prediction using the testing set and also calculate the various classification key performance indicators such as accuracy, confusion matrix and the classification report showing the precision, recall and f-1 score.