AMBEONE TRAINING INSTITUTE

FINAL PROJECT SUBMISSION

EMPLOYEE ATTRITION EDA AND CONCLUSION USING MACHINE LEARNING

(Case Study using the IBM HR employee attrition and performance dataset)

Sajay Sam Oommen

th August 2020

INDEX

INTRODUCTION :

Employee recruitment and retention are two of the most critical functions handled by an HR team for an organisation. Its a normal function to be faced with employee throughout the organization. This is considered a normal function and depending from industry to industry , there are percentages considered as normal attrition.

The task of replacement of a highly trained employee both in the work functions as well as the company culture is a very difficult task. Due to this, companies have started to realise to importance of ensuring the employee is comfortable in his work and interactions within the organization and also in his personal life through many work- life balance schemes. HR teams try to ensure they are in tune with employee needs and work towards retention plans for their employees.

One of the additional tools which has been added to help HR teams and the management team members is Data analysis and machine learning tools. These help to identify trends which would otherwise go unnoticed or wouldnt ger the attention they deserve.

The ability of data to provide a measured verifiable and highly accurate analysis gives an initial path towards improvements in the organisation.

PROBLEM STATEMENT :

We will be using the IBM HR Attrition Data set for conducting Machine learning analysis to identify factors affecting attrition. IBM HR Attrition dataset is a IBM company hosted dataset of employee details within an organisation showing a measured set of active and resigned employees and a key list of details regarding them.

TOOLS CONSIDERED :

There will be two forms of analysis done to the data set

1. EDA through Data visualisation using
    - Matplotlib
    - Plotly
    - Seaborn
2. Classification and Machine learning to identify trends using
    - Scikit Learn – Logistics Regression
    - Scikit learn – Support Vector Machines
    - Scikit Learn – Decision Trees
    - Scikit Learn – K nearest Neighbors
    - Scikit Learn – Naive Bayes
    - Scikit Learn – Ensemble Method

References : Kaggle.com / IBM HR Analytics & Performance

PROBLEM STATEMENT

1. We need to find out and is the key factors leading to attrition
2. What is the likelihood that an active employee will resign from the organization

TYPE OF ANALYSIS

This is a supervised learning task within Machine Learning , we will be considering classification tools.

EXPLORATION ;

There is 1470 rows and 35 columns in the dataset
code : main.shape # to see the number of rows and columns

Age	EducationField	JobLevel	Over18	TotalWorkingYears
Attrition	EmployeeCount	JobRole	OverTime	TrainingTimesLastYear
BusinessTravel	EmployeeNumber	JobSatisfaction	PercentSalaryHike	WorkLifeBalance
DailyRate	EnvironmentSatisfaction	MaritalStatus	PerformanceRating	YearsAtCompany
Department	Gender	MonthlyIncome	RelationshipSatisfaction	YearsInCurrentRole
DistanceFromHome	HourlyRate	MonthlyRate	StandardHours	YearsSinceLastPromotion
Education	JobInvolvement	NumCompaniesWorked	StockOptionLevel	YearsWithCurrManager

- Creating copy of dataset

code : copy1 = main.copy() # making a copy of the file for analysis purpose

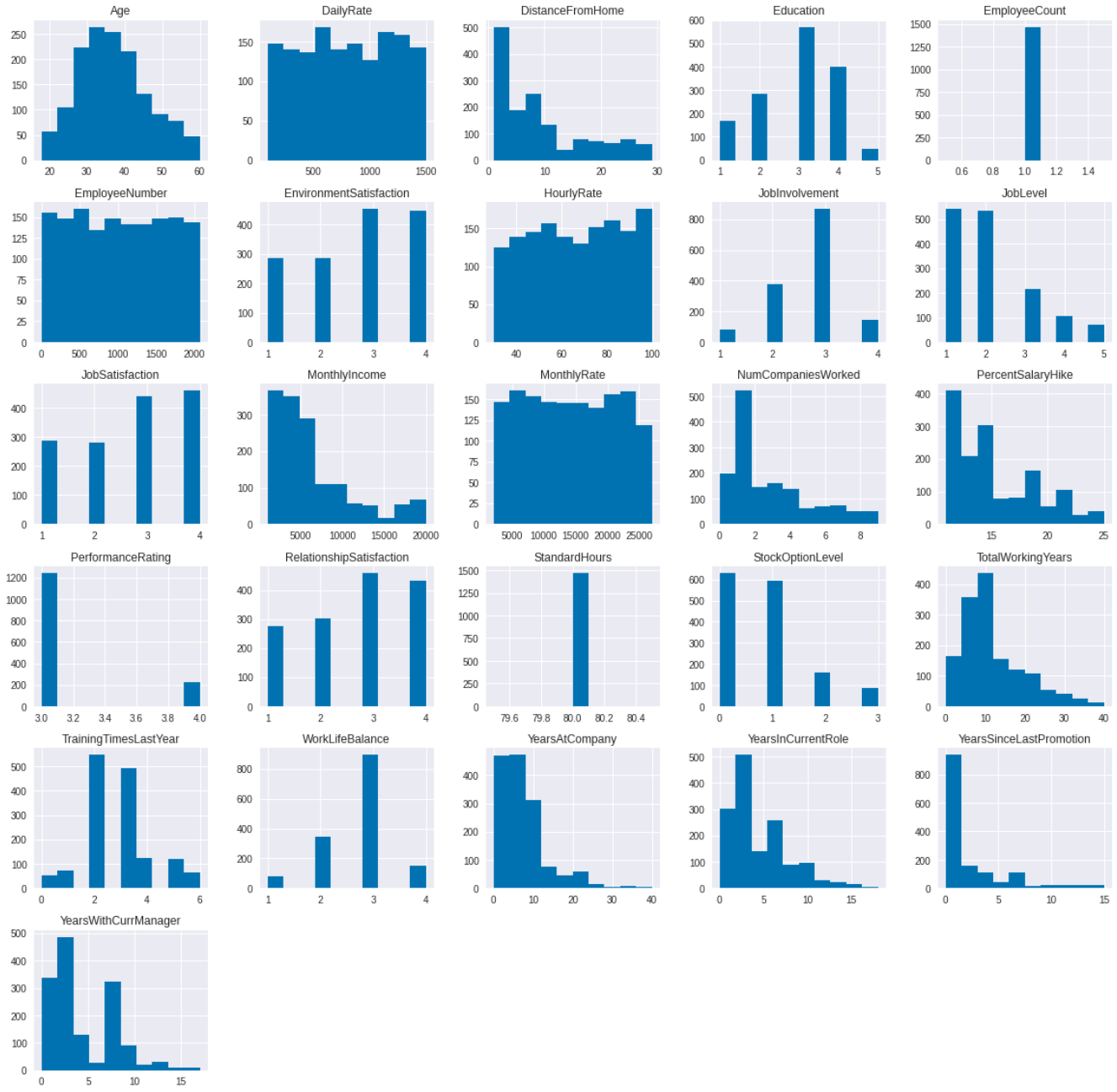
- Finding the Head details of table – identifying the first 5 items in the table to understand the data

code : copy1.head()

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7

- Creating a histogram of all the columns

Code : copy1.hist(figsize=(20,20))

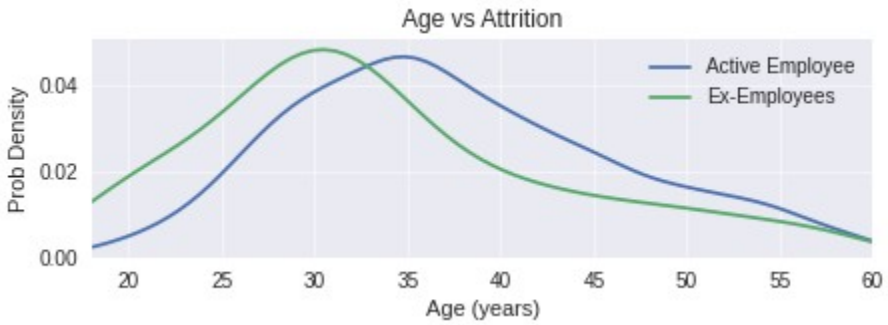


INITIAL CONCLUSIONS:
Histogram helps us to identify the skew of the dataset, eg:
- Age though slight right skewed still shows similiarities to a normal distribution . Approx 25-45
- Also, we identify that Employee Count , Standard hours appear to have no trend or skew and can separated from the analysis

EXPLORATORY DATA ANALYSIS

We will use KDE (Kernel Density Estimation . KDE is a statistical tool used to visualise the shape of a data. It helps to generate a smooth curve for a set of data. This is used for random variables. It is measured on the basis of considering probability density on the y axis.

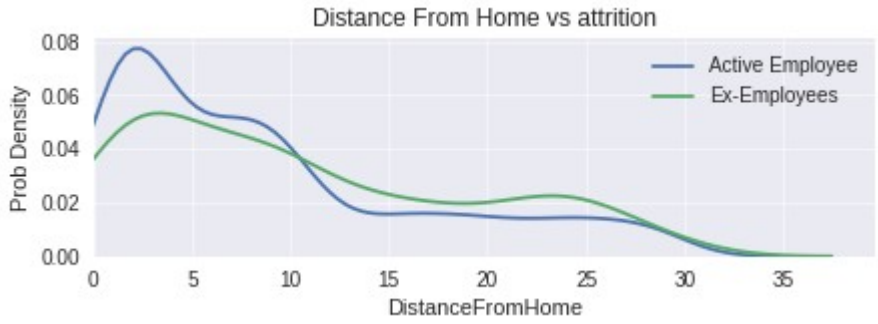
AGE :



Segregated as active and resigned employees.

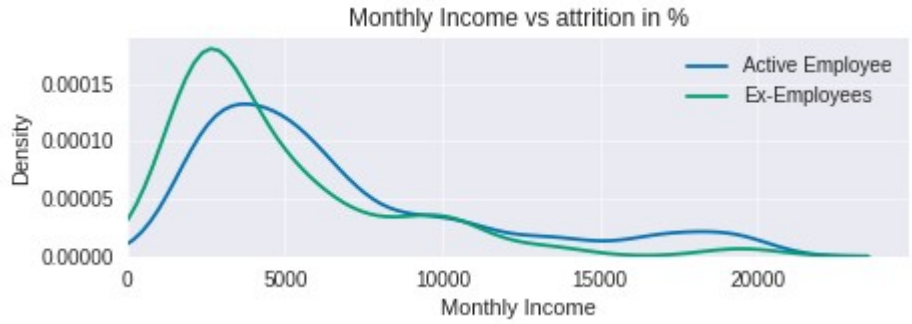
Peaks at 28-30 years which should be the highest age group for ex employees and 3-36 for active employees.

DISTANCE FROM HOME



Average distance from home for current employees is 8.9 kms and ex-employees is 10.6 kms

SALARY AND INCOME



Employee Hourly Rate is between from $30 to $100.

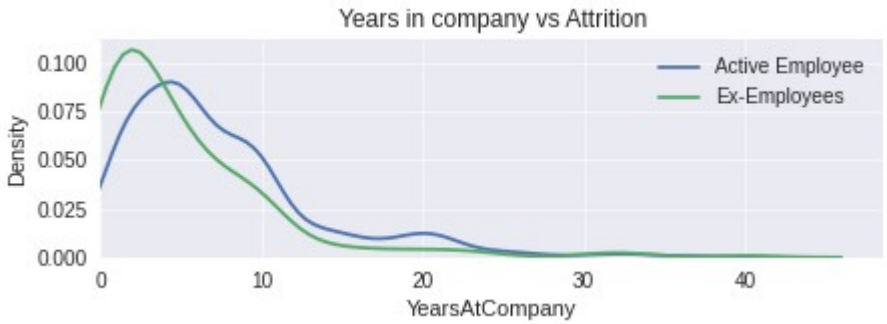
Employee Daily Rate is between $102 to $1499.

Employee Monthly Rate is between from $2094 to $26999.

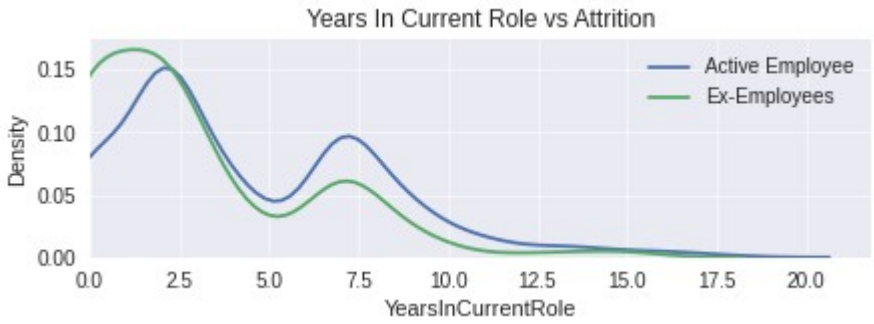
STANDARD HOURS

Standard hours for all is 80 hours.

YEARS AT WORK COMPARISONS

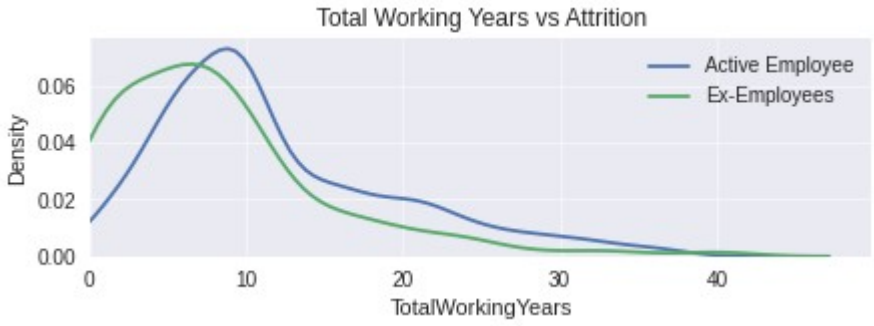


Number of Years is between 0 to 40 years.



Number of Years in current role is between 0 to 18 years.

Years since last promotion is between 0 to 15 years



total working years is between 0 and 40 years

number of Years with current manager is between 0 to 17 years.

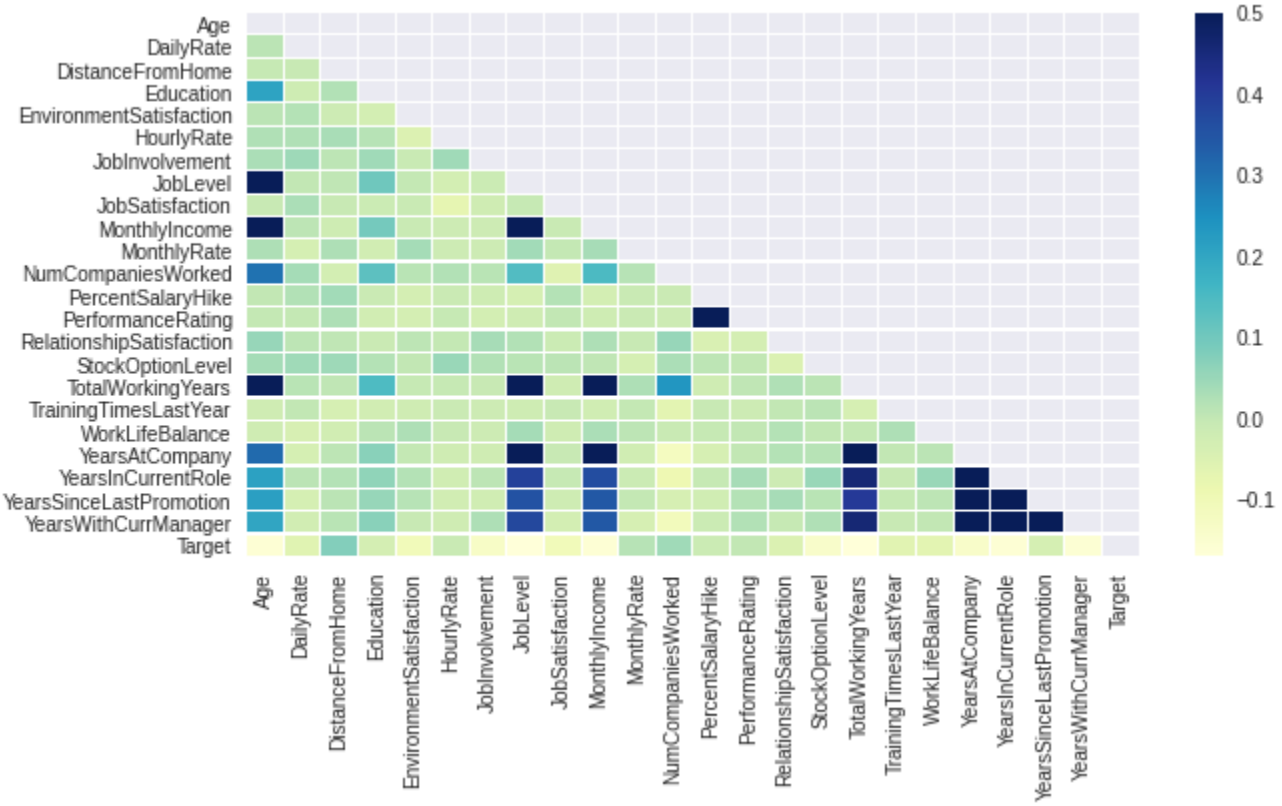
TRAINING TIME LAST YEAR

Number of training times is between 0 to 6 years.

SALARY HIKE

Salary hikes are between 11 and 25%

FINDING CORRELATION BETWEEN ATTRITION AND ALL OTHER FACTORS



CONCLUSIONS

1. Highest negative correlation is between total working years, job level, years in current role
2. Highest positive correlation is between performacne rating, number of companies and monthly rate
3. majority of people left as soon as they completed 2 years after which the trend reduces
4. the longer a person is left in same role, higher is the chances of him leaving.
5. attrition is higher in those who work for overtime.
6. Single employees are leaving in higher numbers than married and divorced.
7. people who travel are higher leavers than others.
8.

ENCODING FOR DATA PROCESSING

Encoding is the process of labeling categorical columns into numberical values.
ALl analysis requires numerical values in order to work. The process of converting encoding helps in this.
Main processes are LabelEncoding and OneHotEncoding.
#ONeHotEncoding is a linear algebra function where a table is created to represent the variable with one column for each category and row for each example.

Code : from sklearn.preprocessing import LabelEncoder, OneHotEncoder

```
# Create a label encoder object
le = LabelEncoder()
```

Label encoding is used for columns with 2 or less unique variables, for the rest of them we have to convert them to dummies.

```
# convert rest of categorical variable into dummy
copy1 = pd.get_dummies(copy1, drop_first=True)
```

	Age	Attrition	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate
0	41	1	1102	1	2	1	1	2	0	94
1	49	0	279	8	1	1	2	3	1	61
2	37	1	1373	2	2	1	4	4	1	92
3	33	0	1392	3	4	1	5	4	0	56

Contd…

FEATURE SCALING

#Feature scaling is used to minimise the wide range in between values within a column and to bring them down
to ensure that machine learning algorithms will be able to analyse them. within a fixed range.

	Age	Attrition	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate
0	2.738095	1	3.579098	0	1.25	0	0	1.666667	0	4.571429
1	3.690476	0	0.633 5	1.25	0	0	0.002419	3.333333	5	2.214286
2	2.261905	1	4.549034	0.178571	1.25	0	0.007257	5	5	4.428571
3	1.785714	0	4.617037	0.357143	3.75	0	0.009676	5	0	1.857143
4	1.071429	0	1.750179	0.178571	0	0	0.014514			

Database after feature scaling.

SPLIT INTO TRAIN AND TEST DATA

Since we have class imbalance (i.e. more employees with turnover=0 than turnover=1)
let's use stratify=y to maintain the same ratio as in the training dataset when splitting the dataset
code :

```
X_train, X_test, y_train, y_test = train_test_split(copy1,
                                                    target,
                                                    test_size=0.25,
                                                    random_state=7,
                                                    stratify=target)
print("Number transactions X_train dataset: ", X_train.shape)
print("Number transactions y_train dataset: ", y_train.shape)
print("Number transactions X_test dataset: ", X_test.shape)
print("Number transactions y_test dataset: ", y_test.shape)
```

BUILDING MACHINE LEARNING MODELS

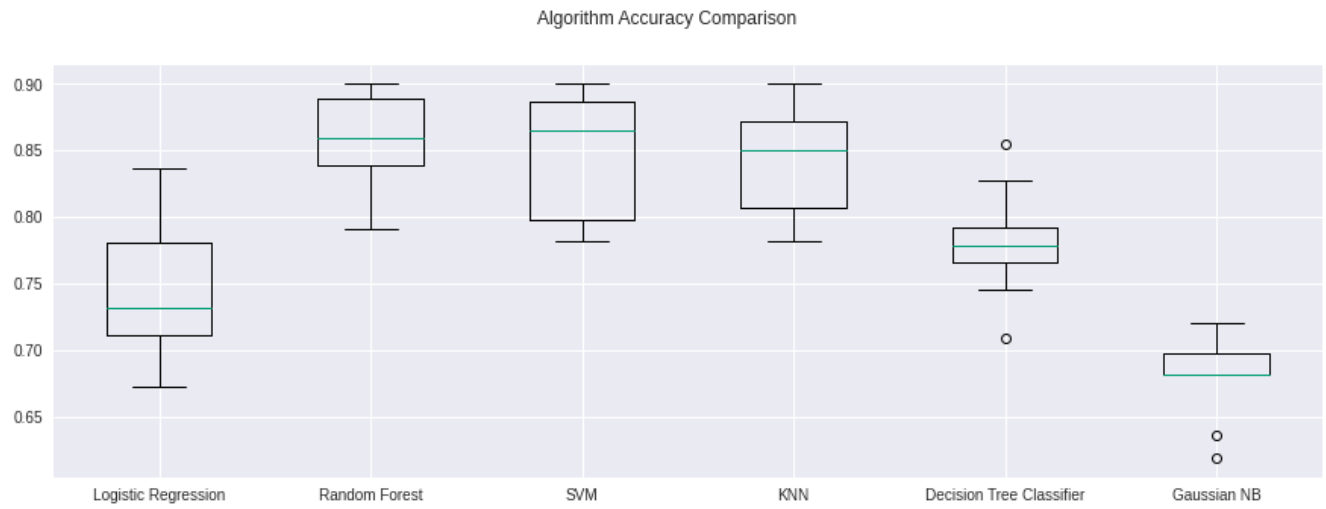
```
# selection of algorithms to consider and set performance measure
models = []
models.append(('Logistic Regression', LogisticRegression(solver='liblinear', random_state=7,
                                                         class_weight='balanced')))
models.append(('Random Forest', RandomForestClassifier(
    n_estimators=100, random_state=7)))
models.append(('SVM', SVC(gamma='auto', random_state=7)))
models.append(('KNN', KNeighborsClassifier()))
models.append(('Decision Tree Classifier',
               DecisionTreeClassifier(random_state=7)))
models.append(('Gaussian NB', GaussianNB()))
```

EVALUATING EACH MODEL

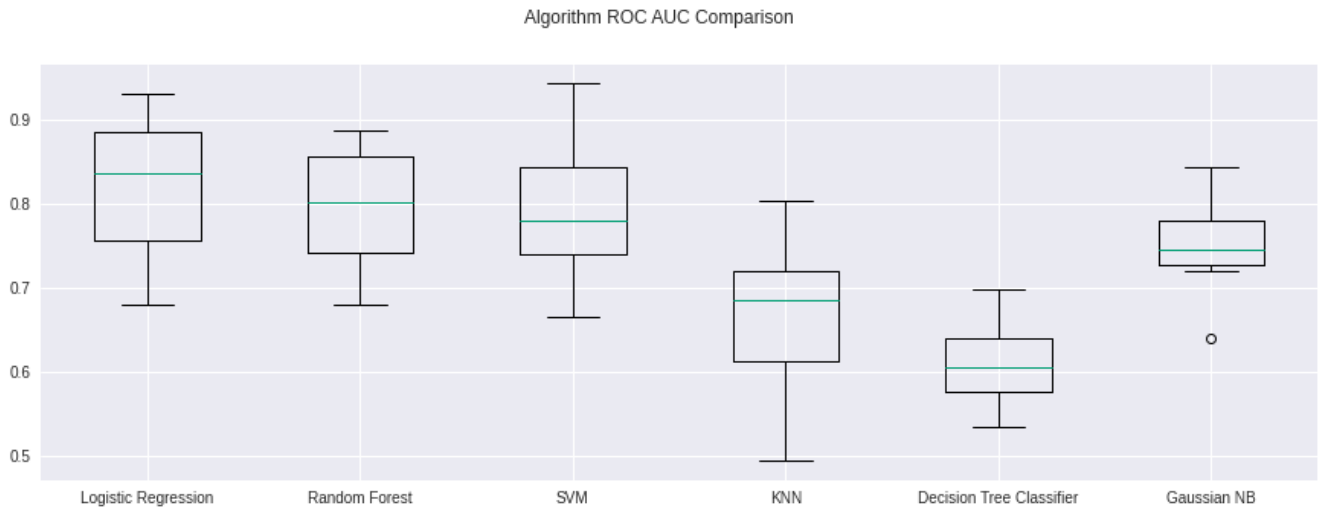
Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD	
0	Logistic Regression	81.82	8.21	74.58	5.52
1	Random Forest	79.41	6.84	85.66	3.71
2	SVM	79.01	8.62	84.66	4.40
5	Gaussian NB	74.96	5.13	68.14	3.09
3	KNN	66.57	8.66	84.39	4.08
4	Decision Tree Classifier	61.24	4.96	78.13	3.82

CLASSIFICATON ACCURACY – Classification accuracy is a study on the numbers of predictions vs total number of predictions.
It is best when the classes have equal number of observations, in this dataset attrition is much lower than active employees.

ALGORITHMS COMPARISONS



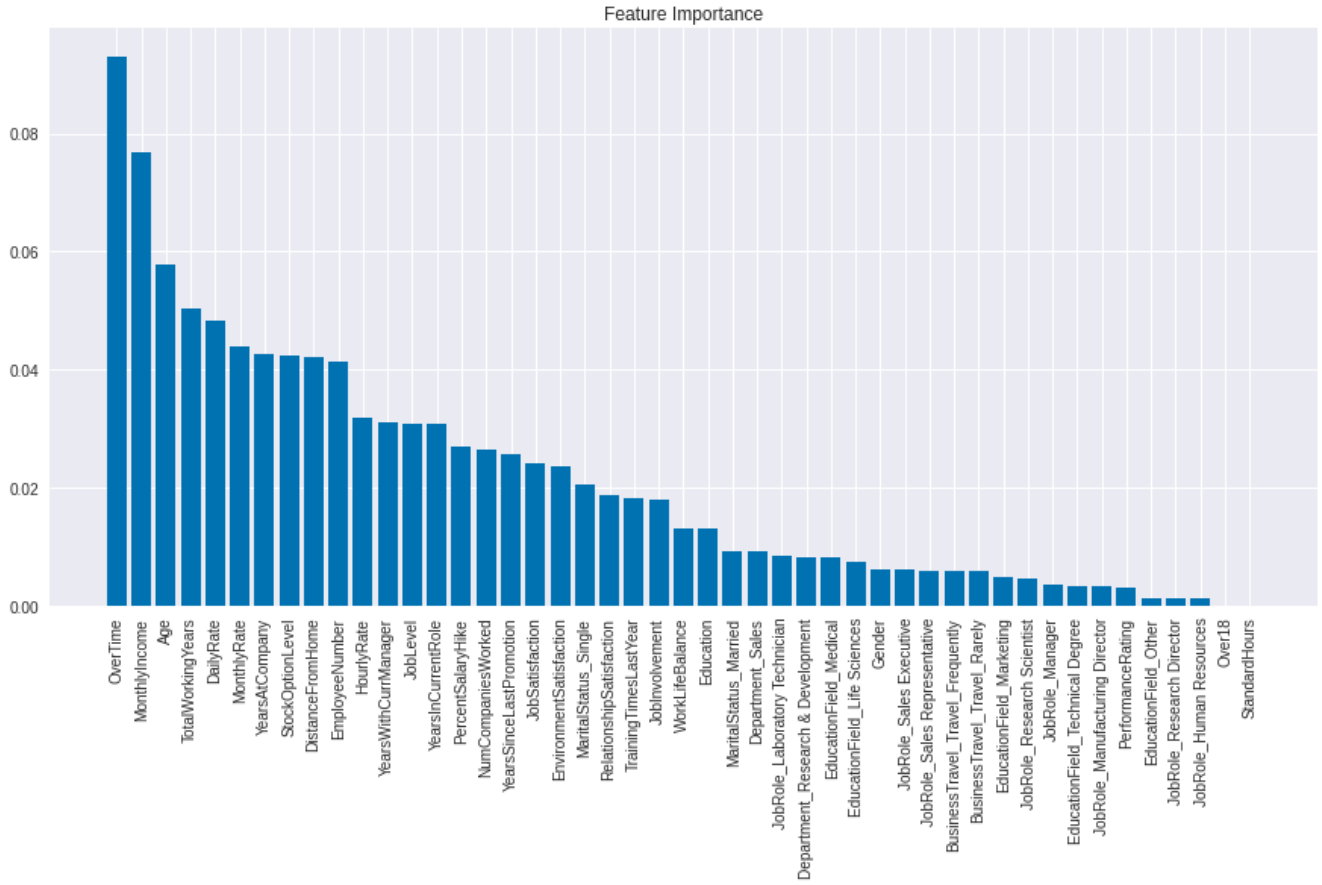
ROC AUC COMPARISONS



Area under ROC curve is for binary classification. This is used to differentiate between positive and negative classes.
On a scale : 1 is perfect accuracy and 0.5 means the model is random.

NOTE : Logistic regression and random forest have shown the most accuracy.

IDENTIFYING MOST IMPORTANT FEATURE THROUGH RANDOM FOREST



Y-Axis scale from 0.00 to 1.00

PROJECT CONCLUSIONS.

WE are considering all the fields which are 0.6 and above.

FIELDS which indicate highest chance of attrition

1. Overtime -Employees with overtime are showing a higher chance of resignation
2. Monthly income – employees on lower income are more chances to resign
3. Age – younger employees have shown a tendency to resign sooner.

PROPOSALS

1. OVERTIME

Inform the managers and their team leaders and work with them to ensure that work flow is improved. Ask them to identify any instances where work is stagnating leading to higher overtime for staff members. This will require higher involvement from the management team to work with the team leaders to identify this and improve it in order to provide employees with a better work-life balance.

2. MONTHLY INCOME

Monthly income can be reviewed internally to check any variations in income within teams. Also we should look at external consultants to provide us with the salary benchmarks for each role to understand how we can align with the market if there are differences.

3. AGE

Younger employees have shown a tendency to resign sooner. We need to check if there are cultural / salary expectations we are not meeting currently. Also understanding the younger staff members through a survey will be helpful to see if there are any changes we can bring in.