Employee Churn Model

Aditya Shekhar Camarushy, Melissa Mathias, Sai Prajwal Reddy
[adcama@iu.edu, melmath@iu.edu, reddysai@iu.edu]

Abstract- A company's churn rate, or employee churn rate, refers to both the attrition rate and the turnover rate. These terms refer to the number of employees who leave the organization during a specified period, usually a year [1]. Of late Employee Churn has become is a very expensive predicament for most organizations considering the resources spent for job-postings, interviews, overtime, sign-on bonus, suboptimal productivity while the new employee acclimatizes .etc. Our project aims to examine the various factors that cause an employee to leave from an organization and predict whether they leave. Using the model we developed companies can reduce the churn rate by taking preemptive measures in changing factors that affect employee churn.

Keywords - Churn, Attrition, Supervised Classification, Exploratory Data Analysis, Categorical Data, Numerical Data, Pipelines, Grid Search, Scaling, One hot encoding, Train Test split, Decision Tree, K Nearest Neighbors, AUC Score, Hyperparameter Tuning, Ensemble model, Random Forest Classifier, XGBoost, Adaboost

I. INTRODUCTION

This project can broadly be classified as a Human resources Management Analytics task with the main stakeholders being the HR department of an organization. As we discussed in the previous section, employee churn can be a very expensive affair in terms of the resources a company has to pour into it when replacing an employee. Based on a study by the center for American Progress [2] it was found that a company has to spend about a fifth of the employee's salary in order to replace them and for higher positions this cost becomes quite prohibitive. With our project we train a classifier to predict whether a person is likely to leave or not given factors such as their Income, Travel, Distance from workplace, Gender, Job Satisfaction, Job Level, Marital status. etc.
We believe our analysis will help the HR understand the various factors that may prompt the employee to leave, and they can possibly come up with preemptive measures to appease the employees, it could be in the form of an increase in pay, reduction in work hours, a change in accommodation with a location closer to work and as a result this would lead to better employee retention and reduced attrition. The dataset we use can be found on Kaggle [3], it is a fictional dataset created by IBM data scientists, it was used to demonstrate the IBM Watson Analytics employee Attrition tool.

II. METHODOLOGY

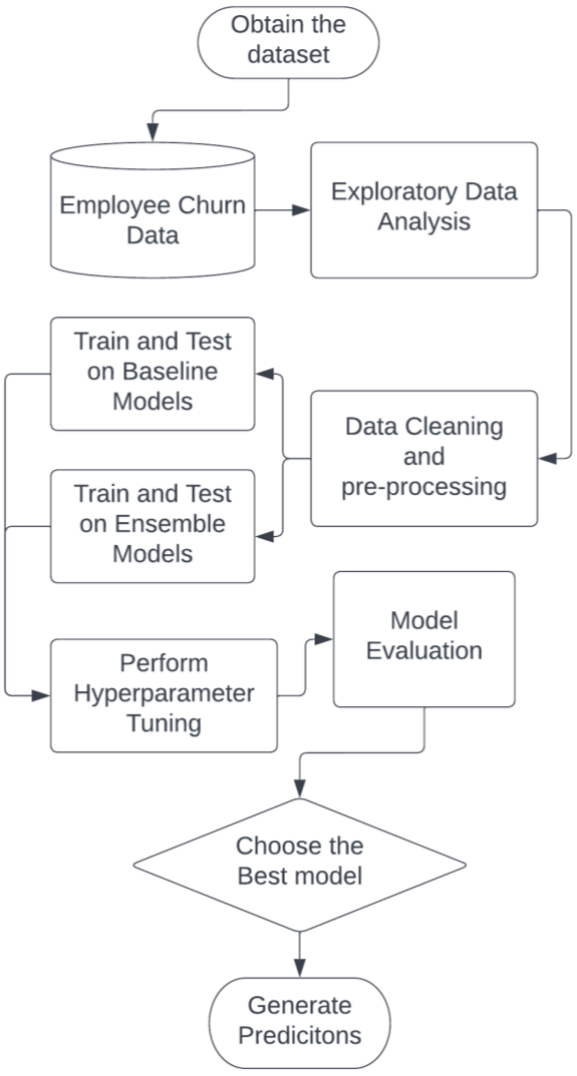

Figure 1. Project Workflow

In this section we will discuss the data, some basic stats on the data, the cleaning and preprocessing steps

following which we will discuss some of the eda results, lastly, we will discuss the models and pipelines we used to perform the classification and predictions.

Data Dictionary

The Dataset consists of one table with 35 columns and 1470 rows, the columns are as follows –

- **Age**: Employee age.
- **Attrition**: The target variable.
- **BusinessTravel**: [Travel_Rarely, Travel_Frequently]
- **DailyRate**: Daily Salary
- **Department:** Employee Department
- **DistanceFromHome**: Distance of home from office in miles.
- **Education**: [1: Below College, 2: College, 3: Bachelor, 4: Master, 5: Doctor]
- **EducationField**: Field of education.
- **EmployeeCount**: A number that is always 1 (Not useful for analysis)
- **EmployeeNumber**: Unique Employee Id
- **EnvironmentSatisfaction**: [1: Low, 2: Medium, 3: High, 4: Very High]
- **Gender**: [Female, Male]
- **HourlyRate**: Hourly Salary
- **JobInvolvement**: [1: Low, 2: Medium, 3: High, 4: Very High]
- **JobLevel**:[1, 2, 3, 4, 5]
- **JobRole**: Employee's designation/ role
- **JobSatisfaction**: [1: Low, 2: Medium, 3: High, 4: Very High]
- **MaritalStatus**: [Single, Married, Divorced]
- **MonthlyIncome**: Monthly CTC
- **MonthlyRate**: Monthly Salary
- **NumCompaniesWorked**:Number of companies worked at
- **Over18**: Flag for >18 or <18 [Y/N]
- **OverTime**: Flag for overtime [Y/N]
- **PercentSalaryHike**: Percentage Annual Salary Hike.
- **PerformanceRating**: [1: Low, 2: Good, 3: Excellent, 4: Outstanding]
- **RelationshipSatisfaction**: [1: Low, 2: Medium, 3: High, 4: Very High]
- **StandardHours**: Hours worked
- **StockOptionLevel**: How much company stock employee owns.
- **TotalWorkingYears**: Number of years in work experience.
- **TrainingTimesLastYear**: Number of hours spent training
- **WorkLifeBalance**: [1: Bad, 2: Good, 3: Better, 4: Best]
- **YearsAtCompany**: Number of years in work experience at the company.
- **YearsInCurrentRole**: Number of years in current role at company
- **YearsSinceLastPromotion**: Number of years since last promotion.
- **YearsWithCurrManager**: Number of years with existing manager.

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	Environm
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	

Figure 2. A snippet of the dataset

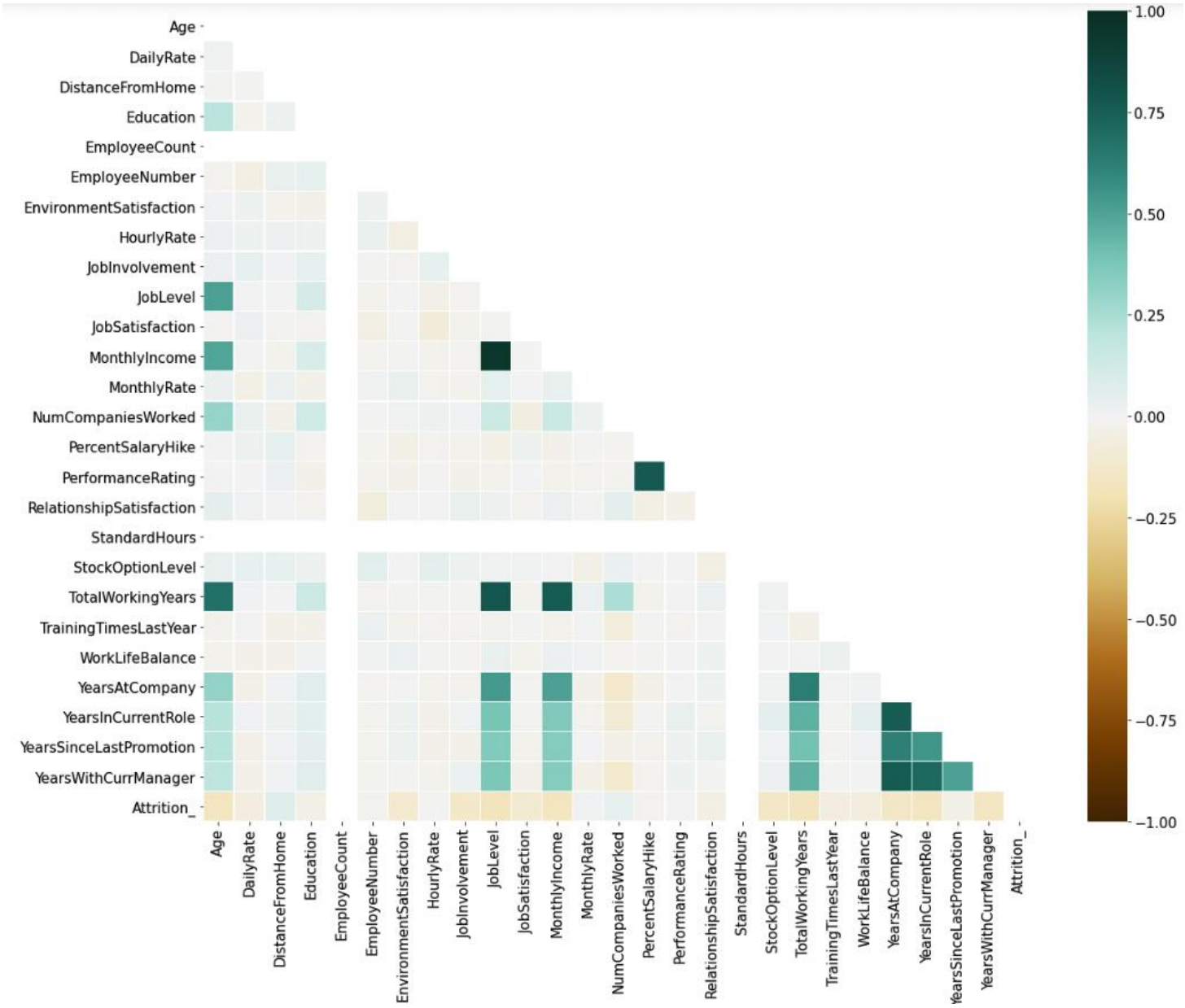

Figure 3. Correlation Plot

Exploratory Data Analysis

We start by performing a correlation analysis to determine the columns that are most important to the analysis. Based on the correlation plot above we can say that Total working years, Job Level, Years in current role, monthly income and age have a relatively high correlation with the attrition rate and they all have an inverse relation, which makes sense. For instance, a person with a higher monthly income would be less likely to quit or a person who is higher in age is less likely to quit as they have responsibilities and so on.

We then proceed to explore the statistical properties of each of the columns in the dataframe that included count, mean standard deviation, min, max, quartiles, uniques, missing count, missing percentage and data types of the column. Another thing to note is that the dataset is imbalanced with 16% belonging to the attrition flag ‘Yes’ and the remaining 84% had a value of ‘No’.

In our dataset we have a few columns, namely Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, RelationshipSatisfaction and WorkLifeBalance have values ranging between 1-5, having particular

descriptions against them as mentioned in the Data Dictionary Section above.

The target/dependent variable is Attrition, as this is a supervised classification problem, 0 indicates active employee and 1 for former employee.

We found that there were 15 categorical columns (Y/N flags and a set of values) and 20 numeric/textual columns. Surprisingly we found that across all columns there was little to no missing data, one plausible explanation for this is that HR Analytics data usually has complete details on all the employee's personal data on-file. It is very unusual for real world data to have no missing values, however we must consider the fact that the dataset we are working with is fictional. We then explore the numerical columns; we found the following-

- The data seems to be quite skewed in the case of Monthly income, salary hikes, distance from home.
- A majority of the employees are in the age ranges of 30 - 45.
- A majority of the employees seem to have worked for the company for 0-25 years.

The next thing we did was analyze at a specific column level by plotting pie charts, barplots, histograms (in the form of a Kernel Density Estimator that normalizes the values across categories and makes comparision easier) as seen in the figures.

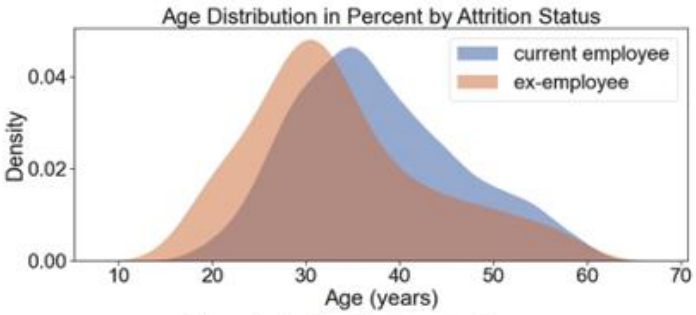

Figure 4. Age Distribution vs Attrition

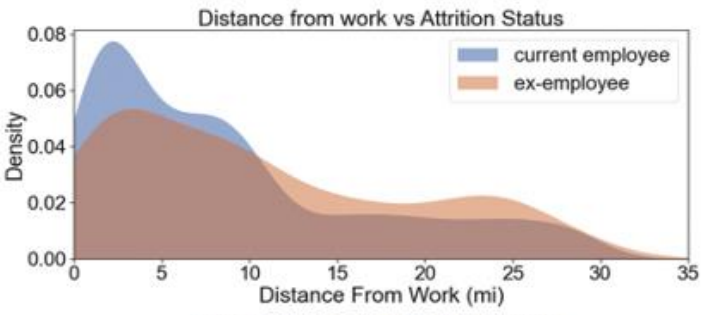

Figure 5. Distance from Work vs Attrition

Fieure 6. Job Level vs Attrition

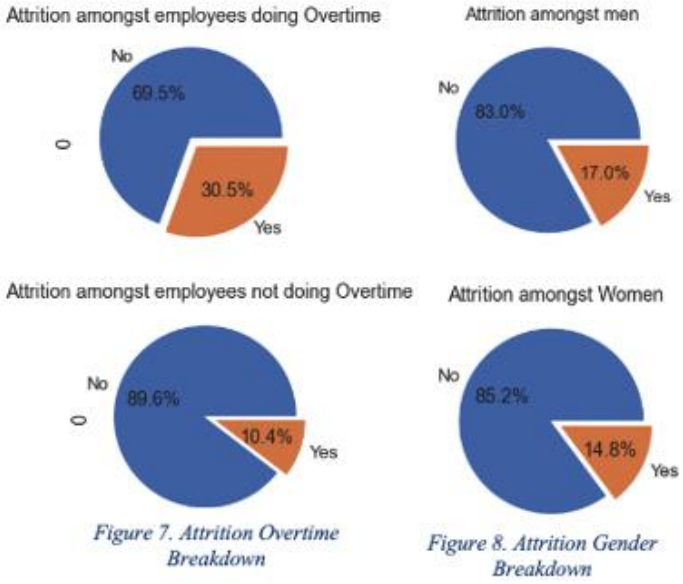

Figure 7. Attrition Overtime Breakdown

Figure 8. Attrition Gender Breakdown

In Fig.4 We see that younger employees are more likely to leave the company as opposed to older employees which logically makes sense.

Fig.5 shows us that employees living further away from the office are more prone to leaving the company which is understandable given the extra time it takes to travel.

Fig.6 shows us that Employees with lower positions tend to leave much more often.

In Fig.7 we observe that employees working overtime are almost thrice as likely to leave.

slightly are more likely to leave as compared to women.

Having more stocks in an organization seems to strongly affect employee attrition, the reason for this is the way stocks vest incentivize people to stay in the company for longer, so offering stocks could be one way to improve employee retention as seen in Fig.9.

In Fig. 10 we see that a good work life Balance vastly improves employee retention.

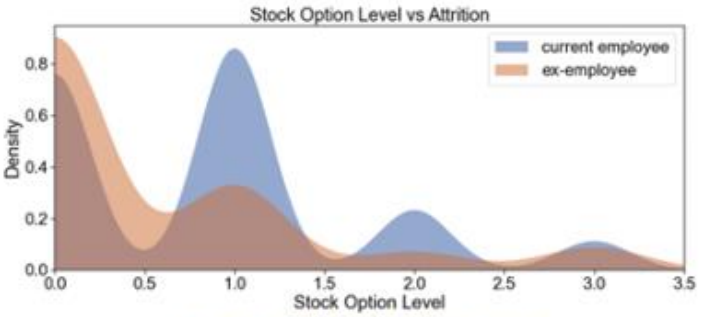

Figure 9. Stock option Level vs Attrition

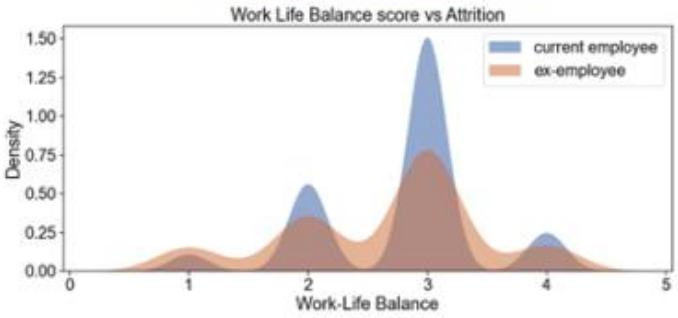

Figure 10. Work Life Balance vs Attrition

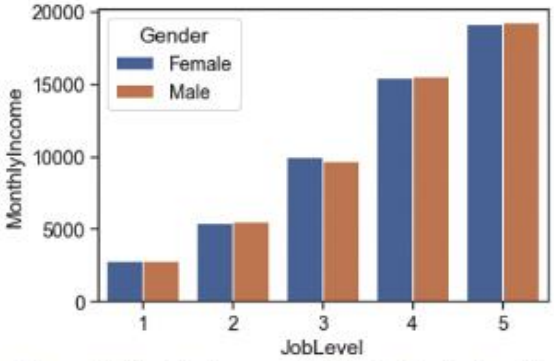

Figure 11. Monthly income vs job level (Gender Breakdown)

Figure 12. Years at the company vs Attrition

Another factor that contributes to a low attrition rate is the longevity of an employee, the longer they are a part of the company the less likely they are to leave as seen in Fig.11.

It is also interesting to note that there is no gender wage gap across different Job levels as seen in Fig.12 however this may boil down to the fact that the dataset used is fictional.

There are still a variety of factors that affect the attrition that we have explored in further detail in our EDA notebook, however we will conclude our EDA takeaways as we have covered most of the important factors that affect Attrition.

Data Cleaning and Preprocessing

Since the data has no missing values and is fictional data there was no necessity to specifically clean, the main thing we focus on here is the data preprocessing.

Initially we extracted numerical and categorical columns into 2 separate lists after which we built 2 different pipelines, one for handling numerical and the other for handling categorical data.
We also performed MinMaxScaling for numerical data and One hot Encoding for categorical data.

The categorical columns are as follows – [Attrition', 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'Over18', 'OverTime'].
The numerical columns are as follows – [Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeNumber', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager']

We then split the data into train and test sets using sklearn's inbuilt train_test_split method with the test data being 30% of the entire data.

We then built a column transformer to selectively apply data preparation transformation pipelines.

Modeling Pipelines

We first built baseline model pipelines with default parameter values, in specific we used KNN and the Decision Tree classifier. We then performed Hyperparameter tuning using grid searchCV in order to get the best parameter values for these two methods.
We then used Ensemble learning methods to further improve the performance of our baseline models, this included Random Forest, Adaboost and Support Vector Machines.

Model evaluation

In order to choose the best model, we need a way to evaluate them, we used Accuracy, Confusion Matrix, AUC scores, AUC / ROC Curves as some of the metrics to evaluate the performance of our models. We will dive deeper into these in our Results section.

> Model Accuracy: Can be is a fraction of number of classifications a model correctly predicts divided by the total number of predictions made.

> Confusion Matrix: It gives us the summary of correct and incorrect predictions with count values and divided down class-wise. It helps us by giving an insight into the errors made by classifier.

> AUC & ROC curves: It shows the performance amongst different classifiers is summarized using the Area under the ROC Curve (AUC).

From the table below Fig.13, we can see that Decision Trees was clearly the most inferior compared to all the models that were used. After hyper parameter tuning there was a small increase in the accuracy and AUC. However, there was room for improvement. We can notice that Support Vector Machines perform the best with an AUC of 0.8505 and an accuracy of 0.8912. We can also notice that most of the models except AdaBoost improved their Test AUC and Test accuracy after hyperparameter tuning while there was a small reduction in the Train AUC and Train accuracy.

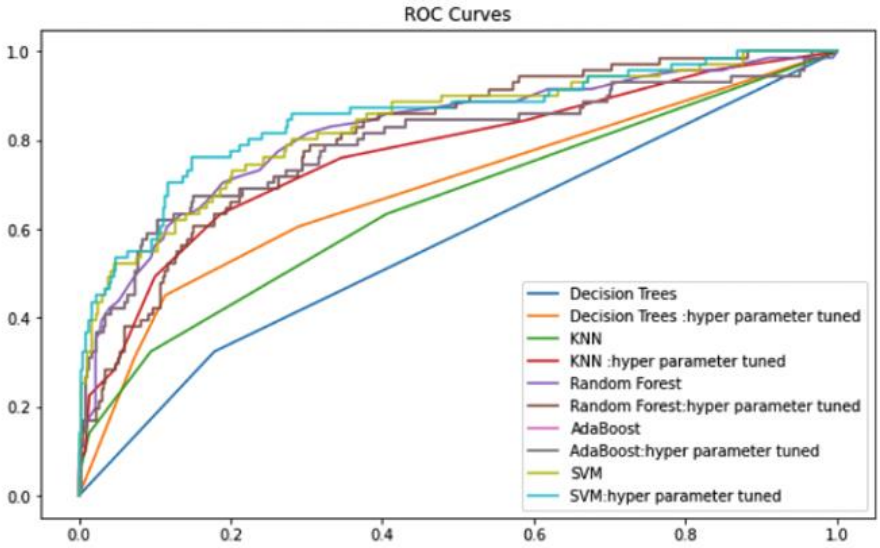

Figure14. ROC Curves for all the models used

We can also notice from the ROC Curve above that Decision Trees do not perform well as the area under the ROC Curve of the model is less compared to other models. We can see that Support Vector Machines and Random Forest perform well as the area under the ROC Curve is very high.

III. RESULT

Model	Train AUC	Test AUC	Train Accuracy	Test Accuracy
Decision Trees	1	0.6080	1	0.7528
Decision Trees with Hyperparameter Tuning	0.7372	0.6902	0.8542	0.8277
KNN	0.8949	0.6546	0.8717	0.8503
KNN with Hyperparameter Tuning	0.8436	0.769	0.8494	0.8413
Random Forest	1	0.8320	1	0.8549
Random Forest with Hyperparameter Tuning	0.9134	0.8003	0.8581	0.8413
AdaBoost	0.9323	0.7947	0.9057	0.8617
AdaBoost with Hyperparameter Tuning	0.9323	0.7947	0.9057	0.8617
Support Vector Machines	0.9587	0.8322	0.8882	0.8526
SVM with Hyperparameter Tuning	0.9099	0.8505	0.9155	0.8912

Figure 13. Result summary

Therefore, the model we would finally select to predict whether or not a person may leave would be the SVM Model in our case, it has a Test accuracy of 89.12 % which is a desirable result, we believe there may be scope for improvement by performing some more advanced feature engineering and implementing advanced techniques like Neural Networks.

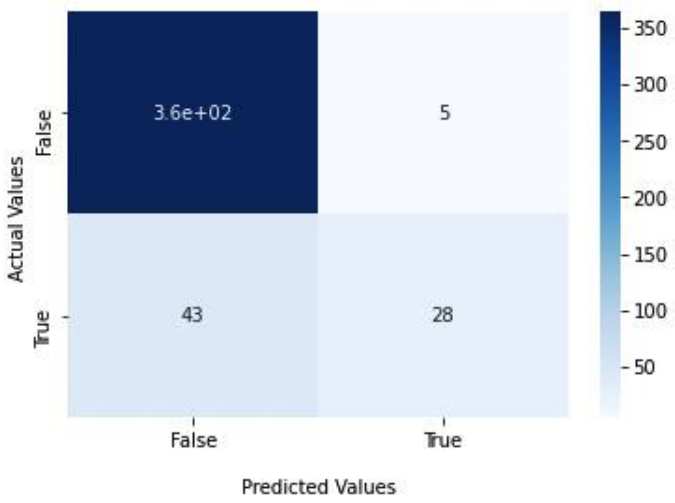

Figure 15. Confusion Matrix for SVM with Hyperparamter tuning

IV. DISCUSSION

The indicators that contribute to higher Attrition (churn) rate are:

1. The Young employees in the age group between 25 and 35 are most likely to switch companies as compared to the older employees. Hence to retain them, companies must adopt ideas like bonus, appraisals, on-site opportunities that leads to clear insight in their advancements, but at the same time even maintain company long term agreement (joining) policies.
2. It was also observed that the attrition rate of employees that live in far-by vicinities is higher than those who stay in a near-by vicinity. As a result, efforts could be made to give assistance in the form of business transportation or Transportation Allowance.
3. Employees with higher salary pay are less inclined to leave the organization. So efforts should be taken to collect information on industry pay standards in the present local market in order to establish if the firm is paying competitive salaries.
4. It was also observed through EDA , that employees who work overtime are more likely to depart. So, efforts must be made to properly plan projects ahead of time, with sufficient support and staff, in order to avoid the issue of overtime and have a work-life balance.
5. In addition to the above factors, it was also noticed that employees who attained stocks of the company do not tend to leave the company, which indicates that employees should be given equity compensation to retain them.

Other steps to reduce employee churn rate that could be taken are:

- Accommodating employees' personal needs, mostly in case of emergencies
- Assuring employees of job security
- Providing an employee-friendly workplace
- Encouragement by managers and higher management for work done, or goals achieved.
- Having a feedback mechanism setup when employee leaves focus on the exit interviews and make any relevant changes.
- Setting manager and employee feedback sessions, to maintain performance along with motivation levels and it indirectly helps in bridging any communication gaps.
- Providing Job dignity and Job security for long term.

V. REFERENCES

[1]https://www.workforcehub.com/glossary/churn-rate/ “Churn Rate”

[2]Heather Boushey and Sarah Jane Glynn. “There Are Significant Business Costs to Replacing Employees ” https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf

[3]Pavan Subash “IBM HR Analytics Employee Attrition & Performance” https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[4] Sisodia, Dilip Singh, Somdutta Vishwakarma, and Abinash Pujahari. "Evaluation of machine learning models for employee churn prediction." *2017 international conference on inventive computing and informatics (icici)*. IEEE, 2017.

[5] Yiğit, İbrahim Onuralp, and Hamed Shourabizadeh. "An approach for predicting employee churn by using data mining." *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2017.

[6] Alamsyah, Andry, and Nisrina Salma. "A comparative study of employee churn prediction model." *2018 4th International Conference on Science and Technology (ICST)*. IEEE, 2018.

[7] Ekawati, Ardhianiswari D. "Predictive Analytics in Employee Churn: A systematic literature review." *Journal of Management Information and Decision Sciences* 22.4 (2019): 387-397.

[8] Dolatabadi, Sepideh Hassankhani, and Farshid Keynia. "Designing of customer and employee churn prediction model based on data mining method and neural predictor." *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2017.

[9] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." *Expert Systems with Applications* 38.3 (2011): 1999-2006.