International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | August 2021 www.irjet.net p-ISSN: 2395-0072

EMPLOYEE ATTRITION PREDICTION USING STACKING AND ITS EVALUATION

Rema V[1], Ruchitha V[2], Mythri K[3], Hariprasad V[4]

[1,2,3,4]*UG Students, Dept. of Information Science & Engineering, BMS Institute of Technology &Management, Bangalore, Karnataka, India*

Abstract— *the recent increase in the technological capacity to gather large magnitude of data and analyze it has changed the way in which decision makers use them to decide on making the optimal decision. Employee attrition very similar to customer churn is an important and deciding factor affecting the revenue and success of the company. To avoid this problem, many companies at the moment are taking guide via machine learning strategies to expect the employee churn/attrition. In this paper we are analyzing data from past and present using different classification like SVM, Random forest (RF), Decision tree (DT), Logistic Regression (LR) and an Ensemble model (EM) to come up with better predictive model for the dataset present. Through this we are hoping to help the company to predict employee churn and take effective measures to retain the employees and improve their economy loss due to the loss of valuable employees.*

Key words— Employee Attrition, **Random Forest, SVM, Decision Tree, Logistic Regression, Ensemble methods.**

I. INTRODUCTION

Organizations have to consider a lot of factors to keep them as a leading company in the competitive market of today. Machine Learning and their techniques have given them a useful tool to get an edge in the market after analyzing data collected over years.

Attrition also called wastage rate or total turnover rate can be considered as a silent killer which destabilizes a company from within. Employees may choose to leave the company for a lot of reasons like equal pay, lack of appreciation, long working hours and many more. As employees are the central source of any company, employee attrition has a negative impact on the revenue of the company along with other various consequences like having to invest more in hiring and training new employees, more pressure on the present employees and a radical decline in the performance of the company. Hence Analytics done using Machine Learning and their tools helps us to understand the issue and source of it, as well as come up with effective solutions for it. Using the past and present data for predicting attrition helps in identifying the causes for the churn and stopping the increasing churn over rate. In our methodology, we have used different classification methods like SVM, Random Forest and Decision tree along with a hybrid model to understand and analyze the performance of different predictive models and compare them using different classification metrics. SVM (Support Vector Machine) are kernel-based algorithms used which serves as a tool to separate different classes. Kernel transforms the input data into higher dimensions where it can be solved using linear classifier by drawing a hyper plane. For example, facial expression recognition is of the uses of SVM where it filters out different expressions into their own class divided by hyper-plane Decision Tree (DT) appears like a tree shaped algorithm to examine and determine a course of action or show statistical probability.

A company may deploy decision trees as a kind of decision support system. Let's consider booking a train to travel as example. First, we look into our calendar to see if a train is available on that date. If available, we look at the time suitable to us. Then we consider the price is within our range etc. Like this at each step we make decisions and go further deep down the branch till an outcome has arrived that is the train being booked.

Random Forest builds a forest with a number of decision tree and is an ensembling method. Logistic Regression uses independent variables for coming to a conclusion. A last method used is Stacking, an ensembling method which combines the predictions from well performing algorithms and gives out a better performance.

II. LITERATURE SURVEY

Bhartiya, N., Jannu, S., Shukla, P., & Chapaneri, R. worked on Employee Attrition Prediction Using Classification Models.
In this study many classification models have been used to predict job attrition using accuracy, area under curve

and confusion matrix as evaluation metrics. This study has shown Random Forest (RF) classifier gives the highest and optimal accuracy at 83.3% and also shows that Naive Bayes (NB) and Support Vector Machine (SVM) are better for classifying (TP) True Positives as shown by greater (AUC) Area under Curve values.

Ray, A. N., & Sanyal, J. worked on "Machine Learning Based Attrition Prediction".
In this experimental study, they have merged probabilistic estimation models with methods like regression and decision trees (DT). This has helped in finding the maximum and minimum attrition group and find out the root cause to solve their problem. Further, initial or base model is refined in accordance to increase the capacity of prediction.

Jain, R., & Nayyar, A worked on "Predicting Employee Attrition using XGBoost Machine Learning Approach".
An XGBoost algorithm is nothing but a tree-based ensemble model working on gradient boosting network. In this paper, they have made use of the high performance of the XGBoost with regards to utilization of memory, accuracy(higher) and running times(low) to predict attrition in the dataset with the objective to help the organization to use this tool to enable them to decrease the rate of attrition.

Brockett, N., Clarke, C., Berlingerio, M., & Dutta, S. worked on "A System for Analysis and Remediation of Attrition".
In this experimental study, they have used an approach known as Clustering for Analysis and Remedial of Attrition (CLARA) and released as an end- to- end system to provide a tool to HR to increase employee retention. A merge or coupling of clustering and scoring measure based on frequent feature patterns in the data of past employees is used to identify and know the present employees who are at a high risk of leaving the company, and then suggest remedial actions to decrease the churn and retain highly valuable employees.

Bindra, H., Sehgal, K., & Jain, R. worked on "Optimization of C5.0 Using Association Rules and Prediction of Employee Attrition".
Having used IBM Watson Human Resource Employee Attrition Dataset, this paper makes a comparative study of prediction of attrition using decision tree and that of it improved with appropriate algorithm. The evaluation metrics used is efficiency in run time and consumption of RAM and is found the decision tree algorithm with appropriate technique performs the former.

Usha.P.M.and Dr. N.V. Balaji worked on "ANALYSING EMPLOYEE ATTRITION USING MACHINE LEARNING".
Data mining is a process through which underlying patterns in the large dataset can be discovered and analyzed. This paper focuses on data mining techniques to understand different factors affecting the employee attrition of human resource in the company or organization using a tool called Weka. Weka, a data science tool used for predictive analytics makes use of algorithms like KNN to cluster data and understand the factors causing the attrition as well as this tool can be used to compare and analyze the performance and effectiveness of various algorithms which the paper makes use of.

Gunjan, V. K., Garcia Diaz, V., Cardona, M., Solanki, V. K., & Sunitha, K. V. N. (Eds.) worked on "Prediction of Employee Attrition and Analyzing Reasons: Using Multi-layer Perceptron in Spark".
Being an open-sourced general purpose cluster framework, Apache Spark is used in Big Data Analytics. This study uses Multi-Layer Perceptron in Spark to predict attrition. The output is analyzed by graphs (plotting them for each attribute and its value of attrition). Also, a user-friendly interface is provided in human understandable language.

Aniket Tambde, Dilip Motwani worked on "Employee Churn Rate Prediction and Performance Using Machine Learning".
In this paper, they have built an expert model to analyze and to predict the rate at which employees are leaving . Machine Learning algorithms, Random Forest and KNN, have been used to predict and Random Forest outperforms the latter having considered confusion matrix as an evaluation method.

Nesreen El-rayes, Michael and Stephen Taylor worked on" An Explicative and Predictive Study of Employee Attrition using Tree-based Models".
In this study, data collected through random resumes in Glass door, a platform for job search, is collected and analyzed to predict the churn of employee due to the job transition. Tree based models like Random Forest (RF) and Gradient Boosting (GB) are used which have given out strong predictive performance when compared to other binary classification techniques.

Mehul Jhaver; Yogesh Gupta; Amit Kumar Mishra worked on" Employee Turnover Prediction System". Taking advantage of the robust nature of the Gradient Boosting technique owing to its regularization nature, it had been used along with three other standard algorithms Support Vector Machine (SVM), Random

Forest (RF) and Logistic Regression (LR) to predict turnover. The output has shown Gradient Boosting technique outperforming the other three algorithms.

III. PROPOSED SYSTEM

1. DATA SET:

A data set is collection of roughly two components. The 2 components are row and columns. Additionally, the main feature here is to consider the records and fields in well-organized structure. For this particular project we have taken employee dataset from IBM which contains 14000 rows and 10 attributes. Each and every detailed information with respect employee is reflected in the dataset used here.

2. DATA PRE-PROCESSING:

We are making use of feature selection method to get features that are important in the given dataset, now let's divide the dataset into 2 essential parts

- Training dataset
- Test dataset

Below are the steps used for dividing dataset:

- Preparing the data in order to get rid of null value or user defined or irrelevant value then separate that whole record from original records and add to dataset which is used for training.
- Once the records are said to be in desired state add it to testing dataset records because it is related to all important features to predict employee attrition.

3. TEST DATASET AND TRAINING DATSET

Categorizing dataset into records used for testing and training the model is the principal part to get accurate outcomes. Due to this process total records deviated and analyzing the behavior of model is easy. In our project we have 9800 records for training the model and 4200 records for testing the model.

DATA CLASSIFICATION TECHNIQUES:

It is a process in which the main goal is to find out in which group data are interlinked together in the provided dataset. Many various ways of classification algorithm are included in our project like Random forest(RF), Decision tree (DT), Support vector machine (SVM), Logistic regression (LR).

1. DECISION TREE ALGORITHM:

The major goal of using this algorithm is to train the model to predict the class with respect to target variables by using its rules. Here comparison starts from root attribute with the record attributes, then jump to corresponding branch to obtain the value else jump to next node.

2. RANDOM FOREST ALGORITHM:

Random forest is the structure generated by combining huge numerous amounts of decision tress that can predict well ruled outcomes. As result each tree here gives an outcome, here these outcomes are evaluated on bases of their weightage.

3. SUPPORT VECTOR MACHINE:

Here the algorithm plots the records in n dimension space with each feature's value, which states value of particular co-ordinate. Mainly used to generate clear margin of separation in high dimensional spaces.

4. LOGISTIC REGRESSION:

Has 2 major variables independent and dependent, its analysis these datasets with respect to data set and obtains a statistical outcome. Examples: For logistic regression can be the force applied on the object to move from rest motion.

ENSEMBLE METHODS:

It is the process of generation of multiple models whose results are combined to improve the outcome of results are combined to improve the outcome of the model. This process of using ensemble generates improved accuracy results. In other words, these multiple models i.e., our base learners define accuracy separately and meta-learner that is our ensemble model combines the process of base-learner to and learn from the best outcomes to improve the accuracy of final models. These ensemble methods are widely classified into:

- Voting
- Stacking
- Bagging
- Boosting

1. VOTING:

Each model makes a prediction for each instance and outcome of models the one with more than half of votes. If it fails to get more than half of the votes, then models are not stable.

2. STACKING:

It is a heterogeneous base-learners which intend to learn in side by side and combine them by making a meta-learner model learn to output a prediction outcome based on the accuracy analysis of multiple base- learners used.

3. BAGGING:

It is a homogeneous base-learners, where learners are trained from sub-datasets parallel and merged into order to improve the accuracy of the outcomes.

4. BOOSTING:

It is a homogenous base-learner, which analyses and works side by side and combines them by making meta-learner to learn to output a prediction based on different base-learners.
Here in our project, we have used the method of stacking in ensemble in order to make sure desired outcome is obtained from the meta-learner after learning from the base-learner.

PREDICTED DATA:

From the obtain dataset we are making sure we avoid the risk of losing employees with more experienced and good work performer by taking necessary step to withhold them in firm.

IV. IMPLEMENTATION

SYSTEM ARCHITECTURE:

The following flowchart represents the overall system architecture. Our first real step is to gather the data, here we have taken employee dataset and then we load our data to perform scrubbing where we normalize all the missing values and clean the dataset.

We will be taking 75% of dataset for training and 25% of dataset for testing. The data pre-processing will be done on training dataset which helps in loading the data into the proper position and preparing it to apply machine learning algorithms.

Once data preparation is completed, we perform data analysis and visualization on different features of the dataset by plotting several graphs. This helps us to get the clear picture about employee features and their distribution and also helps to choose right features during the prediction.

Also, we apply *feature importance* on the dataset which educate us about the importance of the features based on the scores given to them by *feature importance* model. With respect to employee dataset that we have taken, 'satisfaction', 'yearAtcompany' and 'evaluation' are the three features which are best estimator towards the output variable i.e., employee turnover.

Next step is modelling the data, where we train four different ML algorithms: Decision tree (DT), Random forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR). Then we get the prediction of each algorithm.
Now we apply stacked model which is one of the ensemble method to get a hybrid model. Decision tree is used as Meta-model. It is trained on the base model's prediction and on the test data. The training data for meta-model may also include the inputs to the base models.

The training data will fit the meta-model and the final prediction is obtained which is more efficient than the prediction made by individual machine learning algorithm. We are calling this model as "Stacked hybrid model". This Hybrid model not only combine the multiple different skilled machine learning model but also reduces the errors in predictions made by the base models.

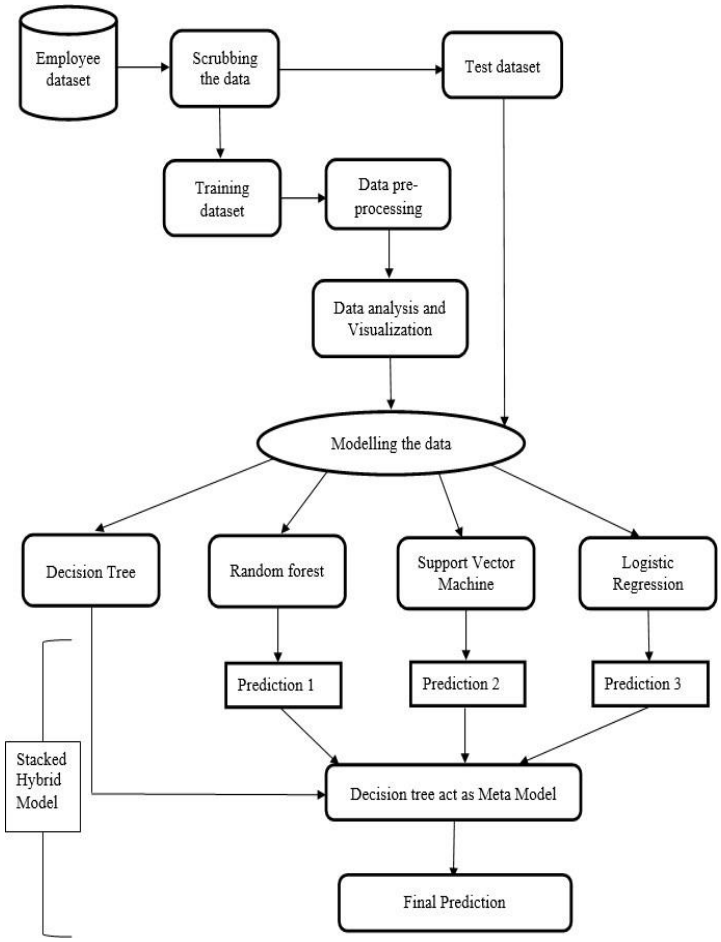

Fig 4.1: System architecture

ALGORITHMS:

ALGORITHM FOR DECISION TREE:
We are making use of ID3 algorithm.
Step1. Find out the information gain of each and every attribute using the following equations:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$\text{Gain(S, A)} = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Step2 If 2 attributes give the same gain value then go to step3, otherwise go to step 4.
Step3. Randomly choose anyone attribute.
Step4. Directly choose attribute with high gain value.
Step5. Make the selected attribute as root node for the decision tree.
Step6. Make selected attribute's value as child node.
Step7. If more unclassified example left then go to step 8, otherwise go to step 9.
Step8. Continue with the remaining attributes. Repeat the process.
Step9. End.

ALGORITHM FOR RANDOM FOREST:

Random forest can be created using 2 different stages:
1. Random forest creation:
Step1. Choose the features randomly such that number of selected features must be less than total features.
Step2. From the selected features, pick a root node 'r' using the best split point
Step3. Make root node's value as child node using best split.
Step4. Repeat the step1 to step3 until single decision tree is formed.
Step5. Repeat step1 to step4 to create many numbers oftrees and construct a forest with that.

2. To make prediction from the random forest created in the first stage:
Step1. Consider the test features and use the rules of every decision tree that has been generated at random to predict and store the outcome.
Step2. For every expected result, determine the votes.
Step3. Find the expected outcome with highest voted as the final prediction from algorithm of random forest.

ALGORITHM FOR SVM:

Step1. Define an optimal hyper-plane which must be maximum margin
Step2. Find the solution for non-linear data as well with the help of kernel method.
Step3. Project data to a high dimensional space where the classification with linear decision surfaces is easier.

Steps to represent an optimal hyper-plane:
Step1. Take the training data of n points:
(X1, y1), (x2, y2)….. (Xi, yi)
Where xi- p-value vector for point 1
Yi- binary class value of 1 or -1
Thus, there are two classes 1 and -1
Step2. Assuming that the data is indeed linearly separable, the classifier hyper-plane is defined as set of points that satisfy the equation

$$\vec{w} \cdot \vec{x} + b = 0$$

Step2. Calculate the hard margin which can be defined as:

$$\mathbf{x}_i \cdot \mathbf{w} + b = 1$$ And
$$\mathbf{x}_i \cdot \mathbf{w} + b = -1.$$

Step3. Calculate the width of hard margin using:

$$2/\|\mathbf{w}\|$$

Where 'w' is the width of the margin.
Step4. Finally, we can find 'w' (weight vector) for the features such that there is a widest margin between two classes.

ALGORITHM FOR LOGISTIC REGRESSION:

Step1. Given a data (x, y), build a randomly initialized matrix for weight. Then, by features, we multiply it.

$$a = w_0 + w_1 x_1 + w_2 x_2 + \ldots w_n x_n$$

Where x- matrix of values
y- vector.
Step2. Pass the output obtained in step1 to a link function

$$\hat{yi} = 1/(1 + e^{-a})$$

Step3. Calculating the cost for this iteration whose formula is as shown below:

$$cost(w) = (-1/m) \sum y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i))$$

Step4. Calculate the derivative of this cost:

$$dw_j = \sum_{i=1}^{i=n} (\hat{y} - y) x_j^i$$

And update the weights:

$$w_i = w_j - (\alpha * dw_j)$$

ALGORITHM FOR STACKED HYBRID MODEL:

There are 3 stages:
1. Build the ensemble:
Step1. Determine the list of base models.
Step2. Determine the meta-model algorithm.
2. Train the ensemble:
Step1. Every 'L' base models are made to learn on the training dataset.

Step2. Perform k-fold cross validation on every base models and collect the cross-validated predictions from each which are represented by p1, p2,….., pL.
Step3. Combine N cross-validated predicted values from every base models to form new N*LN*L feature matrix. The "level-one" data was named in accordance with the original response vector.

$$n\left\{\left[p_1\right]\cdots\left[p_L\right]\right\}\left[y\right] \rightarrow n\left\{\overbrace{\left[\quad Z \quad\right]}^{L}\right\}\left[y\right]$$

Step4. Train the meta-model on the "level-one" data.
Y=f (Z)

3. Predict on new data:
Step1. Generated predictions from the base models will be taken.
Step2. Feed those predictions into the meta-model to generate the final ensemble prediction.

V. EXPERIMENTAL ANALYSIS

In current system only limited amount of techniques are used from the huge collection of data mining techniques for prediction. In above advanced system generated we have applied few algorithms like k-nearest neighbor (KNN), Support vector machine (SVM), Logistic regression (LR), Decision tree (DT) and ensemble model. Fundamentally our data set contain employee, satisfaction grade, assigned projects and work efficiency spent in firm.

In our system we scrub the data so that there are no null values in the data set and if any should remove the null values. We choose the best features for employee attrition and they are given below:

Features	Data type
Job Satisfaction level	Number [10]
Number of Projects	Number [10]
Average Monthly hours	Number [10]
Any Work Accident	Number [10]
Last Evaluation	Number [10]
Time Spent in Company	Number [10]
Department	Varchar [20]
Any Promotions	Number [10]
Salary	Varchar [20]
Turnover	Number [10]

Table 1: Attributes

We use the co-relation matrix and heat map to analysis the below:

CASE 1: Positive (+ve) interdependence
Here project count (PC), average monthly hour (AMH), evaluation i.e., rank of employee is considered. The employee with the ability of covering maximum amount of monthly hours and outputting the completion of multi assigned project with is comparatively high has been observed with highly ranked.

CASE 2: Negative (-ve) Relationships
Here turn over (TO), satisfaction is highly tie-up with each other. In other words, we state that employees having low satisfaction are directly proportional to employees leaving the firm. The heat map is shown below figure 5.1.

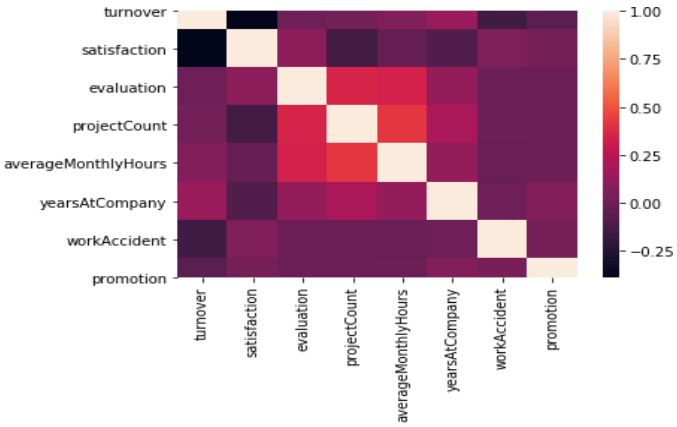

Fig 5.1: Heat map

In order to inspect the scattering on the features. Below are some necessary viewed and hence features:

- **Satisfaction** – Here the employees were categorized into 2 i.e., low satisfaction and high satisfaction.
- **Evaluation** – Using bimodal function performance of employees were categorized into low evaluation (ones below 0.6) and high evaluation (more than 0.8)
- **Average Monthly Hours** - Employees working efficiency is important attribute but able to use a high performance to the fullest is also important so the work time is analyzed with respect to average monthly hour worked by each employee and are categorized again on bases of below 150 hours and more than 250 hours. By this we can say more the working efficiency more will be the average monthly hours worked. These above features are interview with each other.

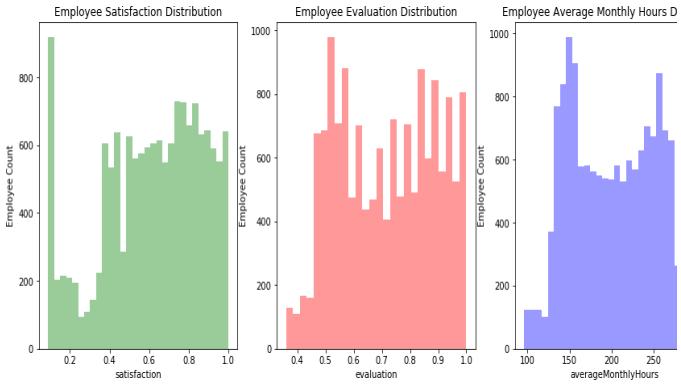

Fig 5.2: Distribution plot

The satisfaction v/s evaluation is the utmost gripping graph. The employees who left the firm were analyzed under 3 main clusters.

- **Cluster 1 (Hard-working and Sad Employee):**

It is important to treat the employees with working efficiency more than 0.75 so that they remain in the firm. Such employees improve the firm and there can someday lead departments in firm. When such employees are not treated well i.e., satisfaction less than
0.2. Such employees may leave the firm.

- **Cluster 2 (Bad and Sad Employee):**

In this case employees with satisfaction level 0.35-0.45 whose working efficiency will be below 0.58. Retaining such moderate employees are necessary as they can be directed by high evaluated to work in order reduce their work load. Hence if their satisfaction is low tendency to leave the firm is more.

- **Cluster 3 (Hard-working and Happy Employee):**

The employees with satisfaction level at between 0.7-1.0 and working efficiency greater than 0.8. Such employees should be treated well in firm in order to keep them associated to firm. They are well satisfied with their work and their performance is highly evaluated.

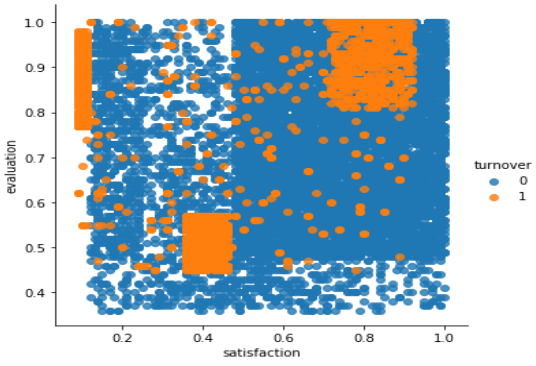

Fig 5.3: Satisfaction v/s Evaluation

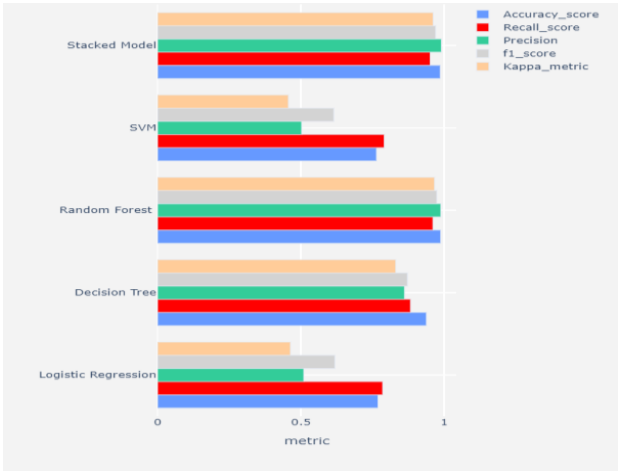

Fig 5.4: Model performances

In the model performances graph we can see stacked model performs the same as random forest which is shown below.

VI. CONCLUSIONS

As we progress with our research, it's obvious that Random Forest along with stacking classifier is the best model to predict employee attrition with an accuracy of 0.98. Clearly, we may say this is a best suited approach. By this we can also conclude by saying, it's important to choose the efficient base models in order to make stacking method more accurate otherwise stacking technique is not recommended.

We trained the computer with limited data (14000 odd records divided into 75percent training data and 25 percent test data), but it can also work effectively on a broad dataset. From the reference above it is very clear how the estimation of attrition plays an important role in thebusinesses.

Workers typically leave when underworked (less than 150hrs / month or 6hrs / day) and overworked (more than 250hrs / month or 10hrs / day), workers with high or low ratings should be taken into account because they are the secret to high turnover levels. Low to medium salaried employees are the most likely to leave the company. Employees who had 2, 6, or 7 project counts were also tend to leave the company. Employee satisfaction is the highest employee turnover indicator. Employee who had 4 and 5 years AtCompany should be considered for high turnover rates Employee satisfaction, yearsAtCompany, and evaluation were the three major factors in determining turnover.

VII. REFERENCES

[1] Qin Zhou," The Impact of Job Satisfaction effect on Turnover Intention: An Empirical Study based on the Circumstances of China", 2017.

[2] Phillips, J. D., "The price tag on turnover" Personnel Journal, vol. 12, 1990, pp. 58-61.

[3] Ms. Ankur Jain, "Impact of TQM on employees' job satisfaction in Indian software industry", 2018.

[4] Boselie, P. and Wiele, T.V.D. (2019) "Employee perceptions of HRM and TQM and its effects on satisfaction and Intention to leave", Managing Service Quality, 2019

[5] Zhu Xiaoyan, Li Yanping," A Study on Psychological Contract, Job Satisfaction and Turnover Intention in Banking Industry", 2018.

[6] Ganesan Santhanam, Raja Jayaraman, Dr.V. Badrinath," Influence of Perceived Job Satisfaction and Its impacts on Employee Retention in Gulf Cooperation Countries", 2019.

[7] Richard, F.G., Joseph, M.L., Billy, B., "Job satisfaction, Life satisfaction and Turnover Intent: Among Food-Service Managers," Cornell Hotel and Restaurant Administration Quarterly, vol. 42(2), 2016.

[8] "Employee Attrition Prediction using Data Mining Techniques", Jeel Sukhadiya, Harshal Kapadia, Prof. Mitchell D'silva, 2017.

[9] "A data mining approach to employee turnover prediction" Amir Mohammad Esmaieeli Sikaroudi1, RouzbehGhousi1, Ali EsmaieeliSikaroudi, 2015.

[10] "Data mining techniques for customer relationship management" Chris Rygielski, Jyun-Cheng Wang b, David C. Yen, 201