**Employee Attrition Prediction**

Machine Learning Project Report

**Team A-01 Members**

Anuja Dixit - [apdixit@uci.edu](mailto:apdixit@uci.edu)

Arpita Sharda - [asharda@uci.edu](mailto:asharda@uci.edu)

Kathleen Sebastian - [sebastk1@uci.edu](mailto:sebastk1@uci.edu)

Matthew Lock - [Matthew.Lock@uci.edu](mailto:Matthew.Lock@uci.edu)

Supriya Shahane - [sshahan1@uci.edu](mailto:sshahan1@uci.edu)

Tatiksha Singh - [tatikshs@uci.edu](mailto:tatikshs@uci.edu)

# Executive Summary

As per Gallup, 51% of the employees in the US are considering a job switch. And around 50-60% of the employees in the US are feeling discontent at work, according to SHRM. According to Future Workplace & Kronos, 87% of the employers have employee retention as a top priority. Losing a high performer employee is costly for a company. Not only the company had already invested in training and on-boarding of the employee but now has to invest in re-training and experience slow work output during the transition.

Our team analyzed an HR dataset with the aim to help companies to solve the employee attrition problem. Data exploration analysis was conducted on the dataset to derive insights and a classification model was recommended to predict whether an employee is likely to leave the company. The report talks about some of the key takeaways based on the analysis. And in the end, some of the limitations/future improvements opportunities are highlighted to make the model more holistic.

# Contents

[Executive Summary](#_heading=h.30j0zll)

[Contents](#_heading=h.3znysh7)

[1. Problem Statement](#_heading=h.2et92p0)

[2. Objective](#_heading=h.tyjcwt)

[3. Dataset Overview](#_heading=h.3dy6vkm)

[4. Data Cleaning/Preparation](#_heading=h.1t3h5sf)

[5. Data Exploration and Insights](#_heading=h.4d34og8)

[6. Testing Approach](#_heading=h.2s8eyo1)

[7. Models Considered](#_heading=h.17dp8vu)

[8. Modeling Results](#_heading=h.3rdcrjn)

[8.1. Benchmark models using Weka](#_heading=h.26in1rg)

[8.2. Models in Weka after preprocessing](#_heading=h.lnxbz9)

[8.3. Models in Python](#_heading=h.35nkun2)

[**ROC Curve**](#_heading=h.1ksv4uv)

[9. Model Selection](#_heading=h.44sinio)

[10. Detailed Description of the Decision Tree](#_heading=h.2jxsxqh)

[11. Key Takeaways/Business Implications](#_heading=h.z337ya)

[12. Limitations/Future Improvements](#_heading=h.3j2qqm3)

[13. Appendix](#_heading=h.1y810tw)

[**Logistic Regression**](#_heading=h.4i7ojhp)

[**Random Forest**](#_heading=h.2xcytpi)

[**K-Nearest Neighbor**](#_heading=h.1ci93xb)

[**Support Vector Machine**](#_heading=h.3whwml4)

[**Lasso Regression**](#_heading=h.qsh70q)

[**Figure 1 - HR Analytics Data Dictionary**](#_heading=h.3as4poj)

[**Figure 2 - Description of the Confusion Matrix for the Dataset**](#_heading=h.1pxezwc)

[**Figure 3 - Decision Tree (Depth 7)**](#_heading=h.49x2ik5)

[**Figure 4 - visuals of all features with the target class (left) in Weka**](#_heading=h.2p2csry)

# Problem Statement

Having a reliable and stable workforce is important for companies to grow and be successful. If employees continually join and leave a company, then tasks that companies need to get done will always change hands. This requires retraining and also ramp-up time. In addition, significant effort is required in the administrative duties to recruitment and administrative matters of onboarding and offboarding employees.

Despite the importance of a reliable employee base, the majority of employees in the US are feeling discontent at work and are considering a job switch. Allied Workforce Mobility Survey discovered that companies lose 25% of all new employees within the first year. All industries are seeing increasing levels of employee turnover, and according to the Center for American Progress, an employee leaving can cost companies 16%-213% of the lost employee’s salary in lost revenue and retraining of new joiners.

Employers worldwide are increasingly making “employee satisfaction” one of their top strategic priorities. However, an investigation is needed to determine the factors that make employees leave and what leads to this “employee satisfaction” that all companies are seeking to improve. Various HR leaders say employee burnout is responsible for up to 50% of their annual workforce turnover, while employees say that the reason, they leave is too many work hours, limited growth opportunities and lack of professional development opportunities.

# Objective

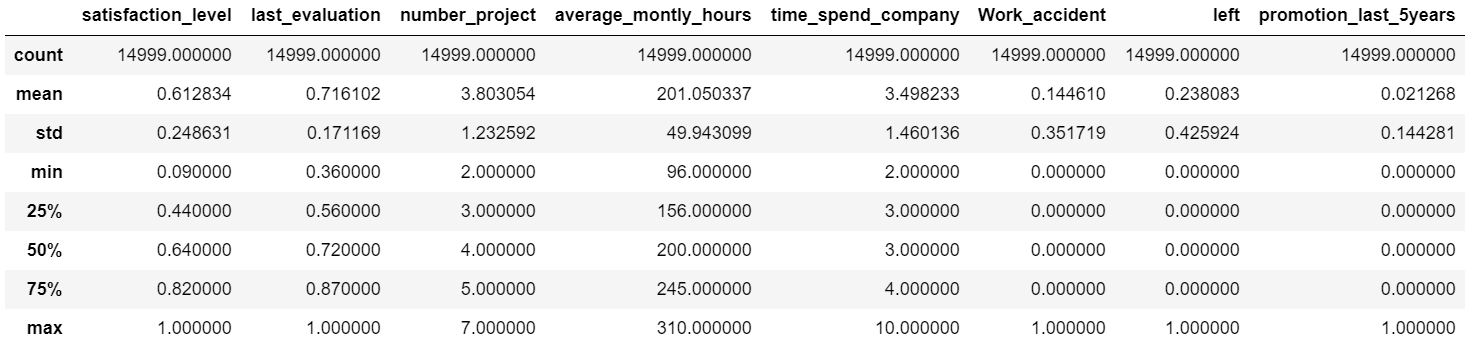
The objectives of this project are to identify the major factors contributing to employee turnover and build a model to predict the employees who are most likely to leave. With the results of our project, companies can foresee employees leaving and create a plan to keep them if the company believes that an employee is valuable, and also alter their business strategy to cater better to those areas that are contributing to the undesirable employee turnover.

# Dataset Overview

Data source: <https://www.kaggle.com/jacksonchou/hr-data-for-analytics>

The HR dataset for this particular project was acquired from Kaggle. It consists of 10 attributes and 14,999 rows. Each row in the dataset represents an employee profile of an anonymous company. For this project, the target attribute is the “left” column, illustrating whether an employee has left a company. Figure 1 (see appendix) depicts a data dictionary that defines each of the attributes used in generating the machine learning models for this project.

The table below illustrates the descriptive statistics for each feature used in the models while also indicating each of their respective statistical measures of central tendency. For instance, the table asserts that attrition (“left” feature) for this particular company has an average of approximately 23.80%, satisfaction level for employees is relatively high with an average 61.28%, and the average time spent for employees at this company averages at approximately 3.5 years.



# Data Cleaning/Preparation

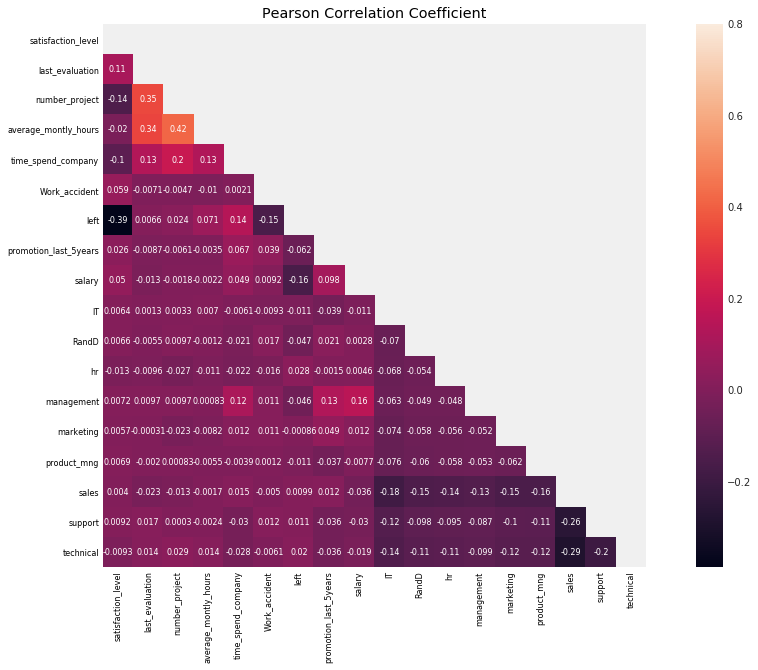
We checked our data for any missing values and couldn’t find any. Our dataset had 3571 records of employees who had left the company and 11428 records of employees who did not. This dataset was then used for data exploration and in Weka models.

Specifically, for models in Python:

* the feature salary (high, low, medium) is an ordinal attribute and was treated by labeling low medium and high as 0,1,2 respectively
* The feature sales (which is actually the various departments that the employees belonged i.e IT, HR, Accounting, Management, Marketing, RD, Product Management, Sales, Support, Technical) is a nominal attribute and was treated by creating dummy attributes (one hot encoding)
* To treat this imbalance in the data, we used the “class” argument= balanced in the decision tree classifier, so that the tree is built on a balanced dataset and the model is not skewed towards the cases of not left (0 - referred as negatives in our analysis) as against the cases of left (1 referred as positive in our analysis).

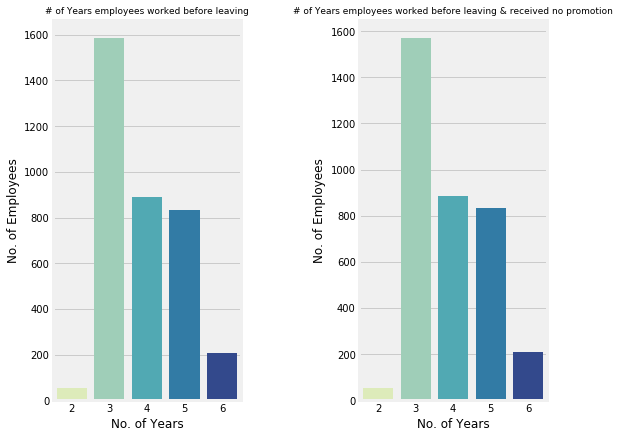
# Data Exploration and Insights

From a total of 14,999 employees, 3571 (23.8%) employees left the company for varying reasons from being generally unsatisfied to being compensated too low. We explore these factors in our preliminary analysis to provide some background information into our project.

We created a Pearson Correlation Coefficient matrix to see what relationship independent attributes had with the dependent attribute ‘Left.’ From the matrix, four attributes contributed the most to employee turnover: Satisfaction Level (-0.39), Salary (-0.16), Work Accidents (-0.15), and Time Spent at Company (0.14). It is unsurprising that these were the top attributes that factored into employees’ decision to leave the company.

Employee Satisfaction is a very subjective term- some may derive satisfaction from having a good work-life balance, while others may have a positive attitude to work if they are compensated fairly, have adequate work to keep them busy, or recognized for their efforts. According to OC Tanner (an employee engagement research company), ‘79% of employees who quit their jobs claim a lack of appreciation was a major reason for leaving.’ It is important for managers to acknowledge their employees’ work and show genuine appreciation so that they know the work and effort they are putting in is not being ignored and taken for granted.

Ultimately, there are many factors that play into employee satisfaction that employers and employees should work together to improve; that way it could potentially ensure valuable employees do not leave.

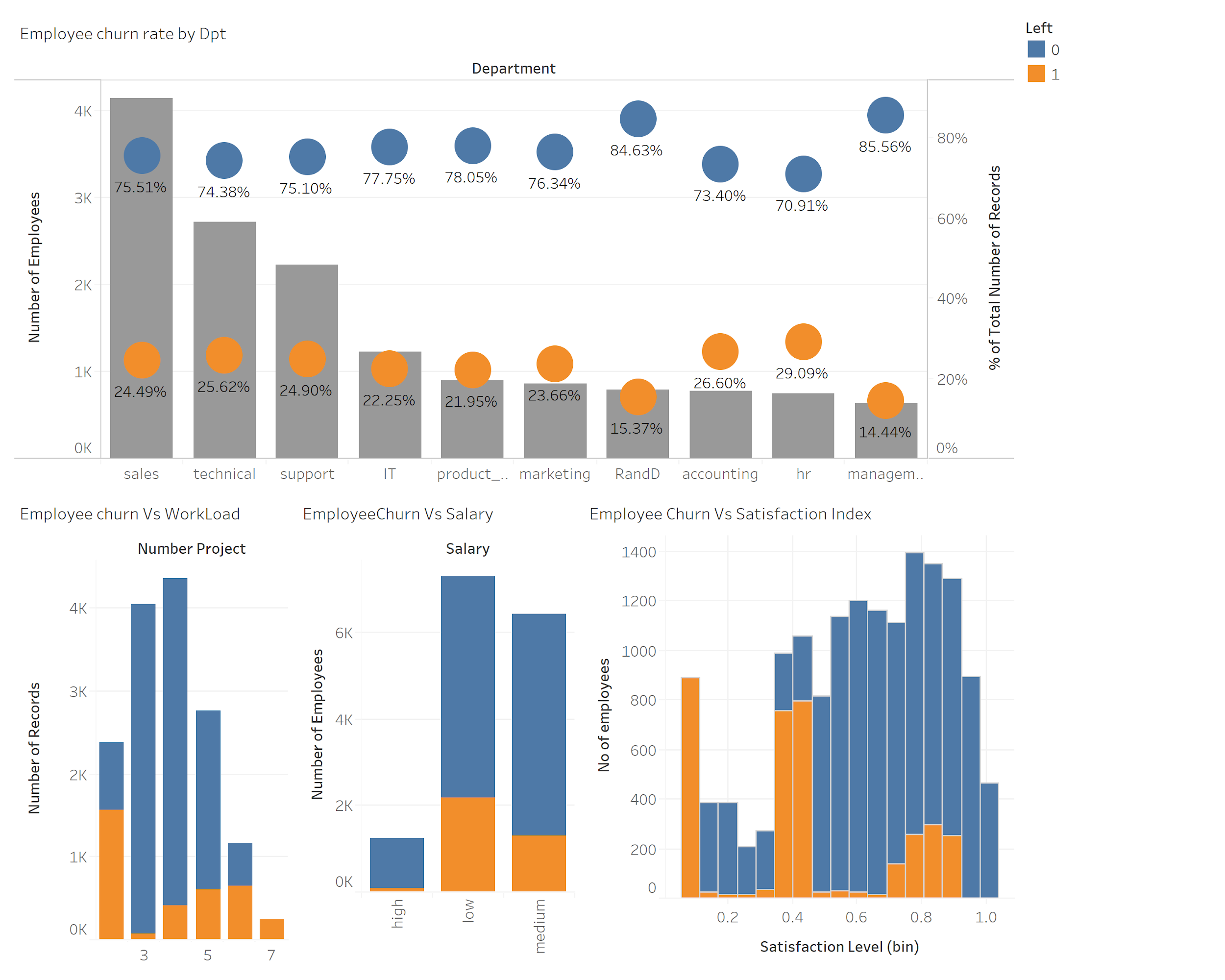


The other attribute that plays an important role in whether employees stay with a company is salary. To many, money is a great motivator and if highly skilled and valuable employees are not being fairly compensated, there is a good chance they will leave a company. Those who consistently perform well, know their value and will use this to their advantage. If they see they are not being fairly rewarded at work, this will oftentimes prompt them to seek new jobs as they feel more confident than other employees in their ability to obtain another job.

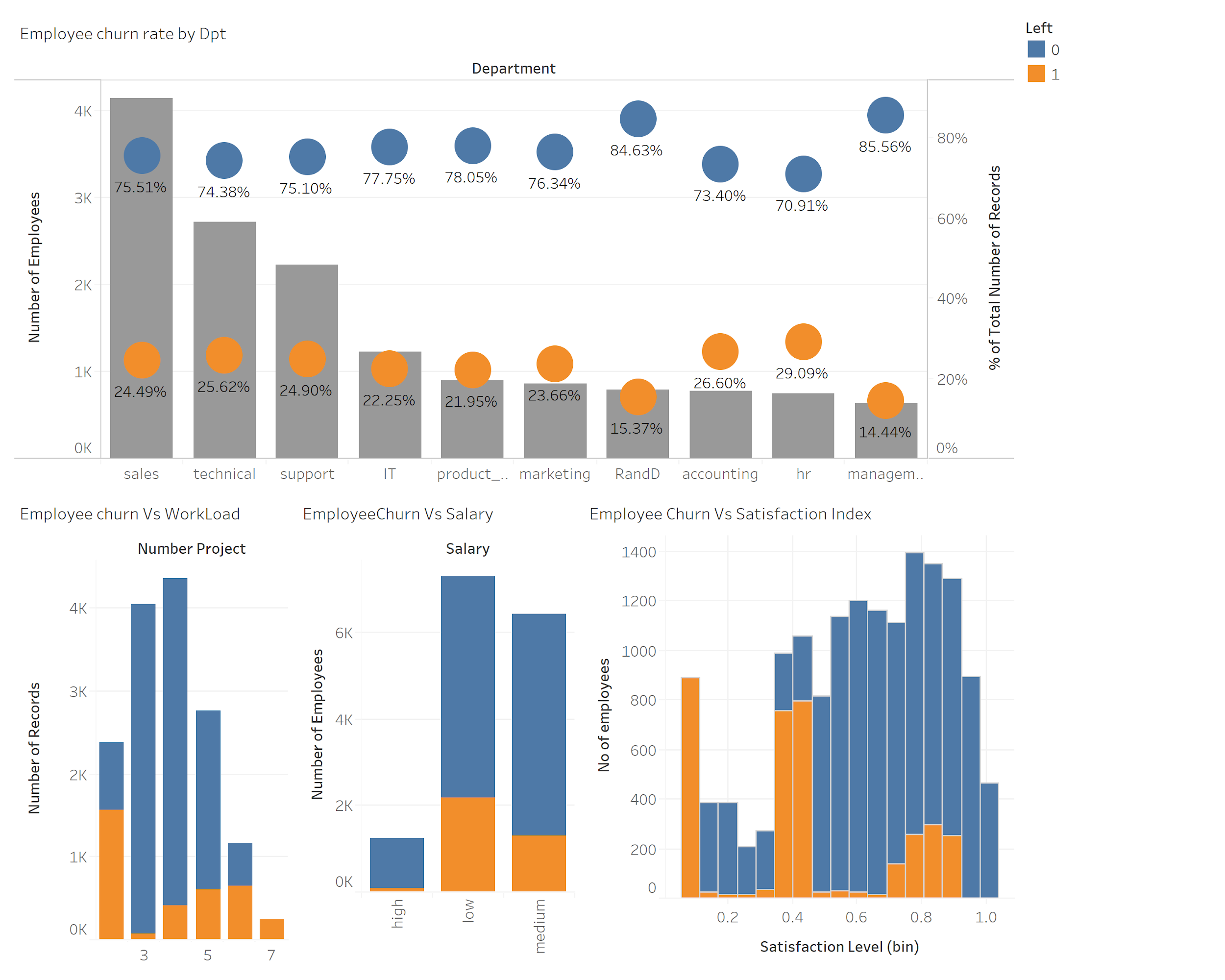
This also ties into promotions at work, however, with our dataset we found that out of all employees who left, almost 45% of employees left after working three years regardless if they received a promotion within the last five years. 99.4% of employees stayed a max of 6 years with the company without receiving promotion. However, those employees who left but received a promotion- which is only 19 (0.005%), stayed with the company a max of 5 years which was interesting.

Work accidents had a low negative correlation which indicates that more employees left without accidents than those who did get into at least one workplace accident; this is also depicted by the bar graph showing the distribution of work accidents. This insight would be more surprising if the company or industry the employees worked for involved some kind of manual labor, for instance- manufacturing, since their safety is sometimes put to risk in these kinds of environments. In this case, we can assume the dataset we obtained contain records from office workers; moreover, this attribute did not attribute to employees leaving the company.

In observing the figure for employee churn rate by department, this visualization depicts the number of employees in each department of the company. Both the sales and technical departments have the highest number of employees, while the HR and management departments have the least number of employees. Overall, employee churn rate across departments varies from approximately 14% to 30%. Despite the high number of records for sales department employees, more than 75% of sales department employees have remained at the company, while only 24.49% of employees in this department left. Another interesting statistic in this graph is that the highest number of employee attrition is actually found within the HR department, with over 26% of employees leaving the company. On the contrary, employees on the management team have the highest percentage of employees who have remained at the company, with over 85% of employees in this department staying. However, it is important to note that these percentages are not an accurate reflection of which departments have the highest number of employees who have left, given that each department contains different number of employees and the percentages will reflect differently based on the number of employees in each department. Provided with these insights, the company may consider focusing their efforts on implementing specific strategies and methods to reduce employee attrition for the sales and technical departments, where there is a high volume of employees and a relatively significant percentage of these employees who have left the company.



Another crucial metric for a company to consider when evaluating employee attrition is to examine the factors that may contribute to employees churning such as workload, salary, and their overall satisfaction level. In this context, satisfaction level is measured using an index that ranges from 0.1-0.5 (0.1 being the lowest satisfaction level and 0.5 being the highest satisfaction level) and is one of the strongest indicators of employee churn. The third visualization depicts a large variability in employee satisfaction level, with more than 55% of ex- employees with a satisfaction level of less than 0.5. The second bar chart pertaining to employee churn and salary illustrates that there is a relatively large fraction of employees with low salary salaries who have churned. Surprisingly, in the first bar chart illustrating employee churn and workload, we also noticed a significant portion of employees who have churned, despite only having a small workload and completing the least number of projects in the company. Given the significant portion of previous employees who have churned, this may be evidence for the company to consider looking into the types of projects they are assigning to employees and analyzing exit interviews or additional historical HR data to determine whether the specific projects being assigned are causing employees to churn.



# Testing Approach

For testing our models, we used the following two approaches:

**Train-test split -**

For modeling in Python, we used the train-test-split from scikit learn’s model selection package. We took 75% data in the training set i.e. 11249 records and the remaining 25% in the test set i.e. 3750 records.

For Modeling in Weka, we used the train-test split in the ratio of 80% training and 20% testing.

**Cross validation -** For modeling in Weka we used 10-fold cross validation. For modeling on python 10-fold cross cross validation in combination with grid search was used for hyper parameter tuning.

# Models Considered

As per the problem statement, we want to predict whether an employee is likely to leave a company or not which means predicting the correct value for the class attribute (left of the dataset. Since this a classification problem, we considered logistic regression, decision tree, random forest, K-nearest neighbor, support vector machine and lasso regression.

We started by creating some benchmark models in Weka. We then proceeded to apply pre-processing steps and run the models again. We also ventured into using Python to create all the models. After reviewing the models, we were able to come to a conclusion on a model appropriate for the business problem which we will cover in the Model Selection section of the report.

# Modeling Results

## Benchmark models using Weka

The dataset .csv file was loaded in Weka and the feature ‘left’ was selected as the target class. We started by running logistic regression and decision tree models.

Since, the overall accuracy as well as other important accuracies like recall, roc curve area, class accuracies are above 90% for the decision tree model, we selected decision tree to proceed with for pre-processing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Logistic Regression** | | **Decision Tree** | |
|  | **Percentage Split**  80%/20% | **Cross Validation**  10-fold | **Percentage Split** 80%/20% | **Cross Validation** 10-fold |
| **Confusion Matrix\*** | a b  2132 149 | a = 0  460 259 | b = 1 | a b  10978 450 | a = 0  2624 947 | b = 1 | a b  2255 26 | a = 0  41 678 | b = 1 | a b  11282 146 | a = 0  149 3422 | b = 1 |
| **Overall Accuracy** | 79.7% | 79.5% | 97.77% | 98.03% |
| **Recall (for class b)** | 36% | 26.5% | 94.30% | 95.80% |
| **ROC curve** | 80.8% | 77.6% | 97.90% | 98.00% |
| **Class a accuracy** | 93.5% | 96.1% | 98.9% | 98.7% |
| **Class b accuracy** | 36.0% | 26.5% | 94.3% | 95.8% |

\* See appendix for a detailed description of the confusion matrix for the dataset.

## Models in Weka after preprocessing

We focused on decision tree to proceed with pre-processing. Since the accuracies for 10-fold cross validation method was more than the percentage split method for testing in the benchmark models, we have selected to run the decision tree model using 10-fold cross validation testing method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Preprocessing steps** | **Overall Accuracy** | **Recall (for the target class b)** | **ROC curve** | **Class a accuracy** | **Class b accuracy** |
| Bins = 10/ Unpruned | 97.03% | 95.00% | 97.40% | 97.6% | 95.0% |
| Bins = 10/ Pruned | 96.55% | 89.90% | 95.90% | 98.6% | 89.9% |
| Bins by freq/ Unpruned | 96.90% | 95.0%% | 97.20% | 97.5% | 95.0% |
| Bins by freq/ Pruned | 96.57% | 90.10% | 95.80% | 98.6% | 90.1% |
| Bins by freq/ Unpruned / instancesperleaf = 20 | 96.31% | 89.80% | 97.90% | 98.4% | 89.8% |
| Bins by freq/ Pruned / instancesperleaf = 20 | 96.21% | 96.10% | 89.10% | 98.4% | 89.1% |

Not only do we want good overall accuracy, recall, roc curve and class accuracies, we also want the tree to be easy to understand. But the decision tree model after pruning was still too complicated to comprehend. Therefore, we proceeded to try out some models in Python.

**Note:** We also ran PCA in Weka which created 17 attributes by using all the features in the dataset. Therefore, we ignored PCA for the modeling.

## Models in Python

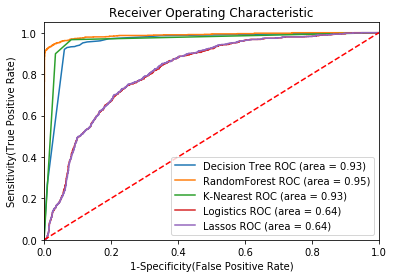
We ran models like Decision Tree, Random forest, Knearest neighbours, Logistic Regression, Lasso Regression on the prepared data in our quest for a model that can give good accuracy, fair interpretability and most importantly immediate implementability. While accuracy of Random forest classifier and K-nearest was better than the decision tree model that we ran, both of them did not satisfy our criteria of easy interpretation and immediate implementation.

Below are the key results of these models (See Appendix for more details on the models).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Scoring Criteria** | **Decision Tree Classifier** | **Logistic Regression** | **Random Forest** | **K Nearest Neighbor** | **Support Vector Machine** | **Lasso Regression** |
| Precision | 92.4% | 61% | 99% | 89% | 55% | 49% |
| Recall | 96.3% | 36% | 91% | 90% | 25% | 25% |
| Accuracy | 97.7% | 79% | 98% | 95% | 77% | 76% |

### **ROC Curve**

We ran a Receiver operating characteristics curve for all these models to see that what accuracy these models give for different thresholds. Random forest has the maximum area under the curve and hence the best performance followed by KNN and decision tree classifiers.



# Model Selection

After creating various models in Weka (benchmark/with pre-processing) and as well as in Python, we noticed that for many models the overall accuracy, recall, class accuracies, roc curve percentages were good (>90%). The question then arises is which model should be selected. If we just consider the overall accuracy, then Random Forest model in Python is the winner. But based on the problem statement and the business use case we are trying to address, we focused on two criteria for model selection:

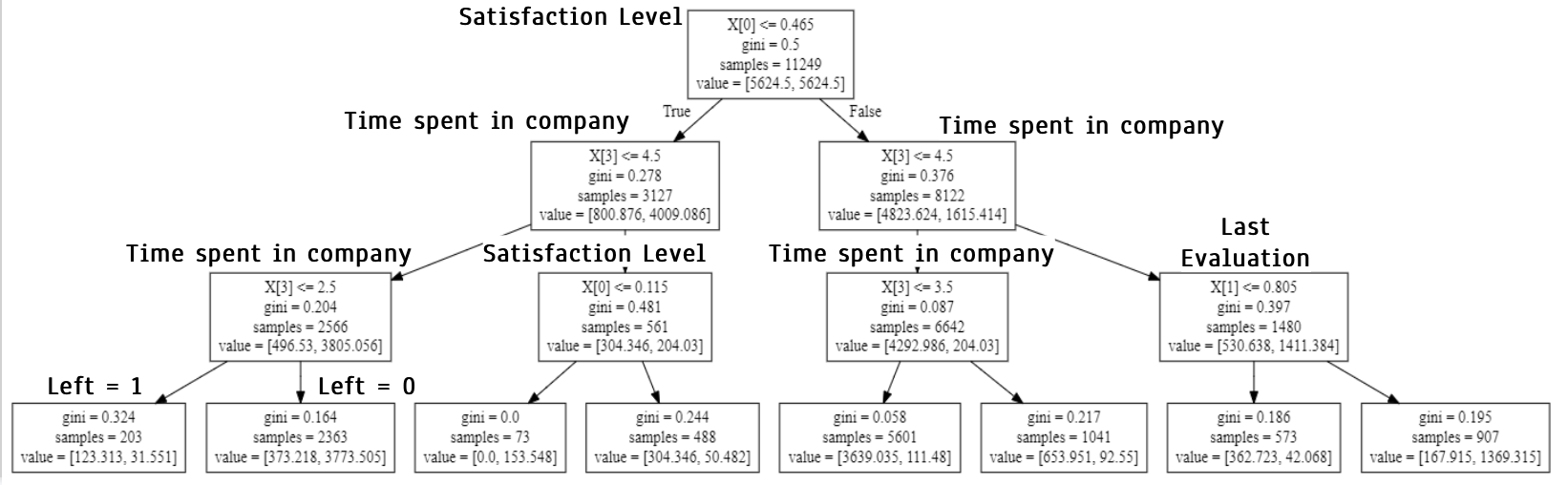
**Recall -** High overall accuracy or roc curve is not enough, the model should also have high class accuracies and high recall (for class b). Consider Recall, we want a high recall percentage because it means that we have less false negatives (see appendix for a detailed description of the confusion matrix).

**Interpretability -** To implement the model in the real world, we want a model that is easy to interpret. Therefore, we were inclining towards decision tree model as it is easy to interpret.

In the end we decided to select the python decision tree model (see the next section for details on the model) for ease of interpretation. Below are some of the reasons why:

* The python decision tree model had only 4 features in the tree as the most statistically important features
* Even though the accuracy reduced, the depth of the tree was reduced to 3 to have a simpler and more pruned tree

**Decision Tree (selected model)**



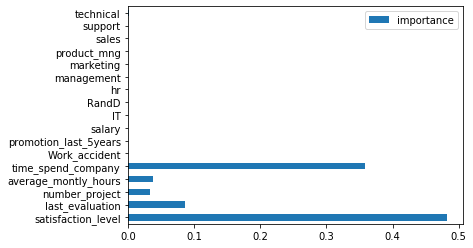
Note: Left = 1 means employee is likely to leave the company and Left = 0 means employee is not likely to leave the company.

# Detailed Description of the Decision Tree

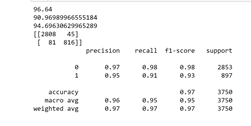
To get better interpretability of the decision tree and to make sure that we get split on the most relevant features we first did hyper parameter tuning. For hyperparameter tuning we used gridsearch CV function from scikit learn library’s model selection package. This function basically runs the model on given values of tree depth and sample size and suggests the combination of parameters that give the best accuracy. We gave the input of depth ranging from 3 to 21. For sample size we gave input ranging from 50 to 500 in steps of 50, and we ran this grid search on a 10-fold cross validation. From this grid search we got the following combination of parameters to give the best accuracy -criterion-”Gini”; Maximum depth of the tree=7 and sample size=50.

We fitted our decision tree classifier with above specified arguments on our training data. Then on this model we ran “feature\_importance” function to identify the most relevant features for an appropriate interpretable split. The 5 most important features that we got from this model in sequence are.

1. Employee Satisfaction level
2. No of years the employee has been in the company (time\_spend\_company)
3. Rating in the last evaluation (last\_evaluation)
4. Average monthly hours an employee spends in the office (average\_monthly\_hours)
5. Number of projects an employee is handling (number\_project)



We ran the decision tree classifier with these 5 features and got an accuracy of 96%, true positive rate or recall of 91% and precision of 95%.



However, the decision tree obtained despite being pruned was still not interpretable (See Appendix for the decision tree of depth 7).

To make it more interpretable we ran a model with a depth of 3 and ran feature importance again to see if the relevant features changed with the changing depth. After running it on the depth of 3 we got the following 3 features as relevant.

1. Employee satisfaction level
2. Rating in last evaluation
3. Years employee has spent in the company

Now we ran a model again on these 3 features and got an accuracy of 91%. Though our accuracy dropped by running model on depth 3 with the 3 most relevant features, still we chose to do that because we wanted to make our model implementable by the business. We did not want to give action items that involved so many attributes that it becomes hard for the human resources to get a control on all of them.

# Key Takeaways/Business Implications

Based on the model, following are some key takeaways:

* Satisfaction Level, Time Spent in Company, Last Evaluation are the major factors to look for employee turnaround
* The decision tree can be implemented in an HR department as a guidance to suggest which types of employees are likely to leave. How to read the tree (few cases):
  + Satisfaction level is the most important feature determining the future of an employee in the company
  + Employees with satisfaction level > 0.465 and time spent in the company > 4.5 years and whose rating from last evaluation is > 0.805 are very less likely to leave the company
  + Employees with satisfaction level < 0.465 and time spent in the company < 2.5 years are most likely to leave the company
* Business could devise employee retention methods focused on high performing employees which are likely to leave the company
* Consider conducting additional analysis for motivations of employees leaving to determine whether employees were laid off/fired or whether it was their choice to leave

# Limitations/Future Improvements

The dataset did not have the following information. If available, it will be very helpful to derive more specific insights and also would be interesting to explore how the various models behave after incorporating the new information.

* Interestingly, salary was not a statistically important feature in the selected model. The reason might be that the salary feature in the dataset is an ordinal feature (low, medium, high values). Would have been good if the data set had the exact salary $ amounts
* Gender information was not available. Having gender information can be very helpful in deriving insights from the data in terms of whether a male or female employee is more likely to leave, or whether a particular department shows more churn in female employee attrition
* Having the information whether an employee left willingly or was fired
* Would be interesting to capture employee sentiment while leaving the company as a categorical feature

# Appendix

### **Logistic Regression**

We ran logistic regression by splitting 70% of our training set and leaving 30% for testing purposes. Without cross validation we got almost 80% accuracy but a very low recall of 36% for our predictor attribute left. We then tried running 10-fold cross validation, however our model still performed the same.

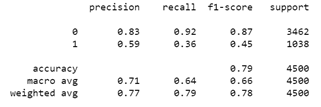
[without cross validation]

Confusion Matrix :

[[3200 262]

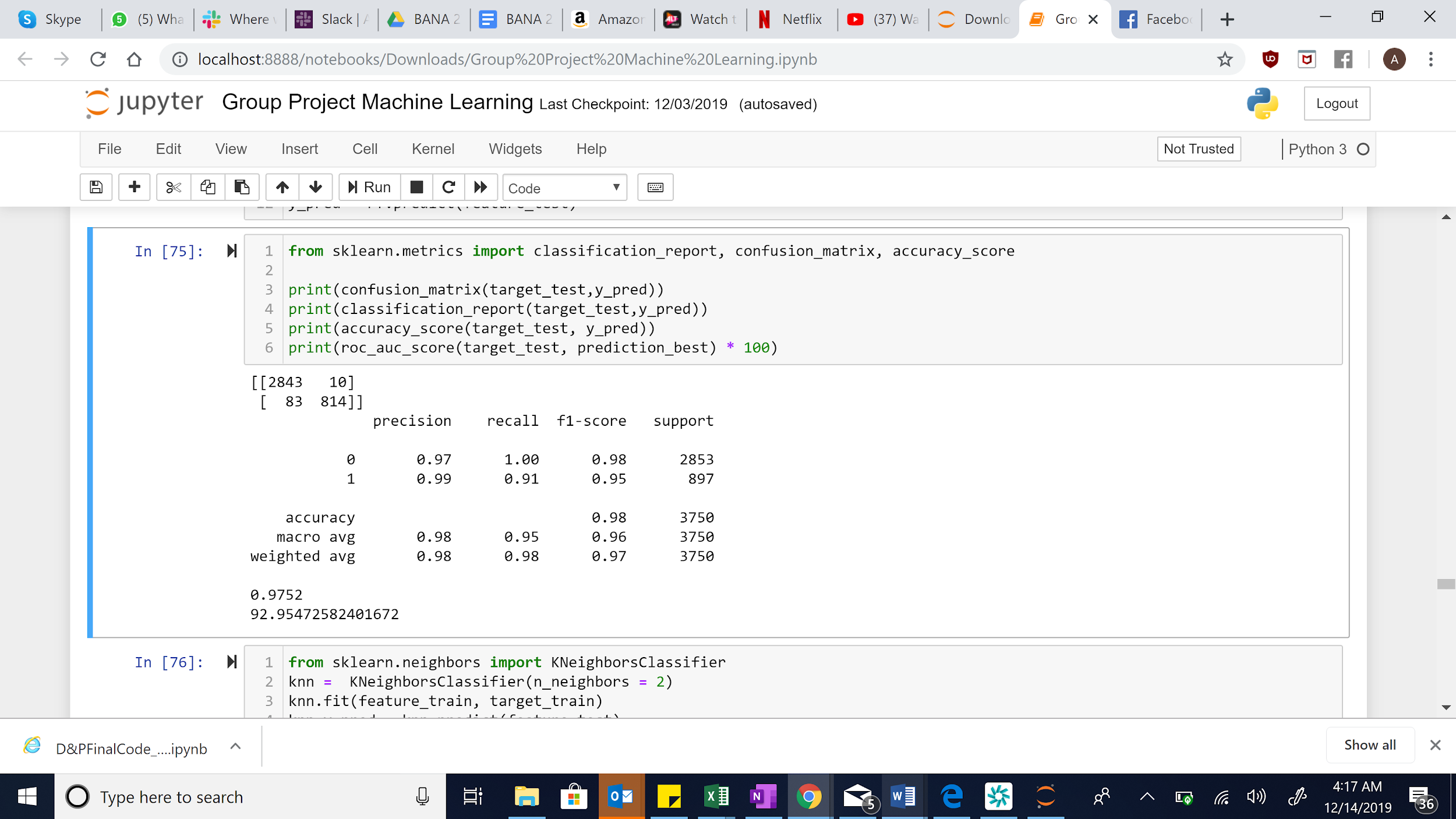
[ 662 376]]

Accuracy : 0.7946



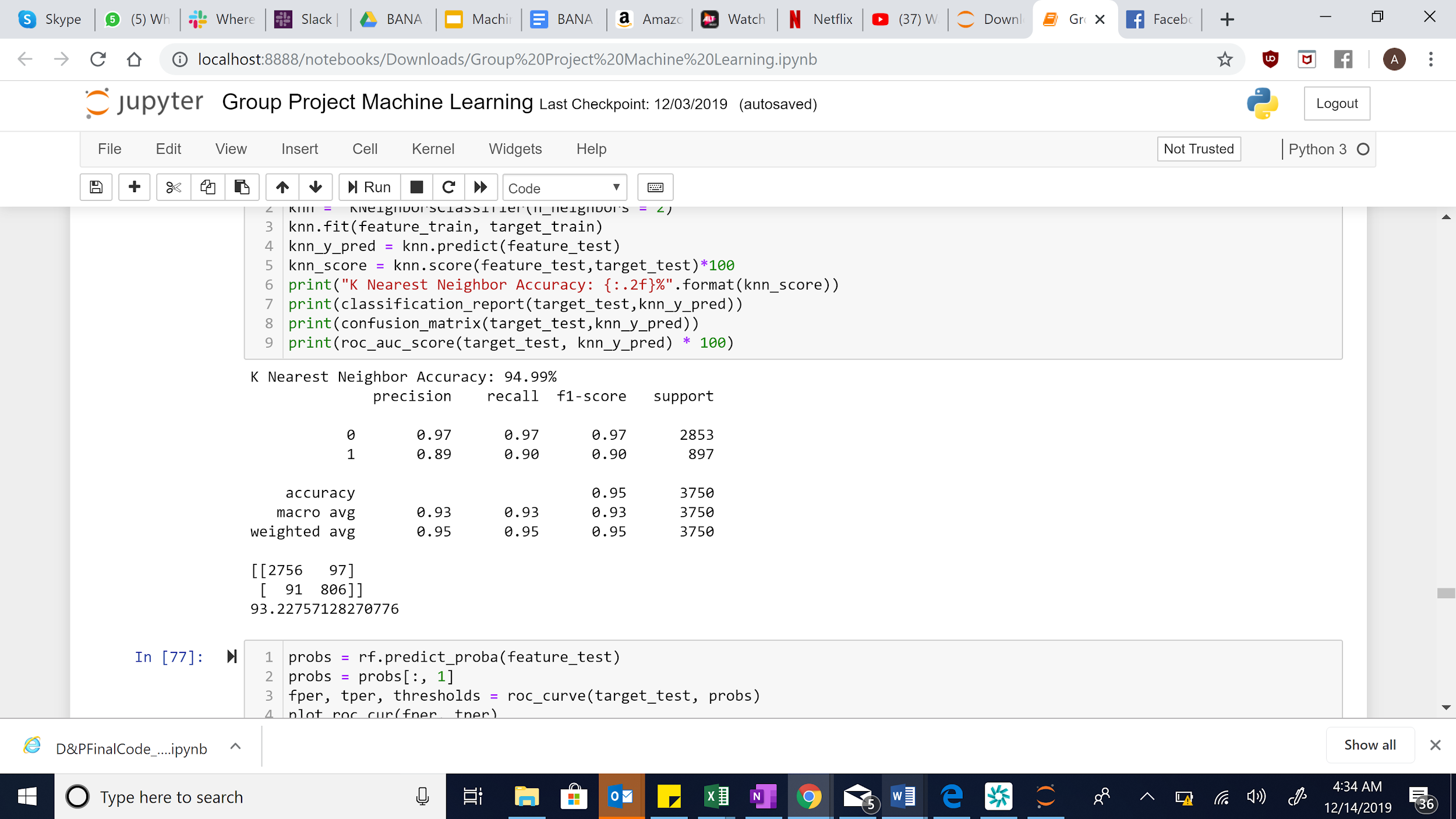
### **Random Forest**

We ran a grid search on random forest classifier for a depth of 4-8;criterion-{Gini, entropy},no of estimators-{200-500}; features-{auto,sqrt,log2} along with a 10-fold cross validation. from this grid search we got the criterion as Gini, depth as 8, feature as auto and no of estimators as 200. With these arguments we ran a random forest classifier and got an accuracy of 97%. However, area under roc curve was still 92% only.(Almost equal to decision tree ROC-AUC)



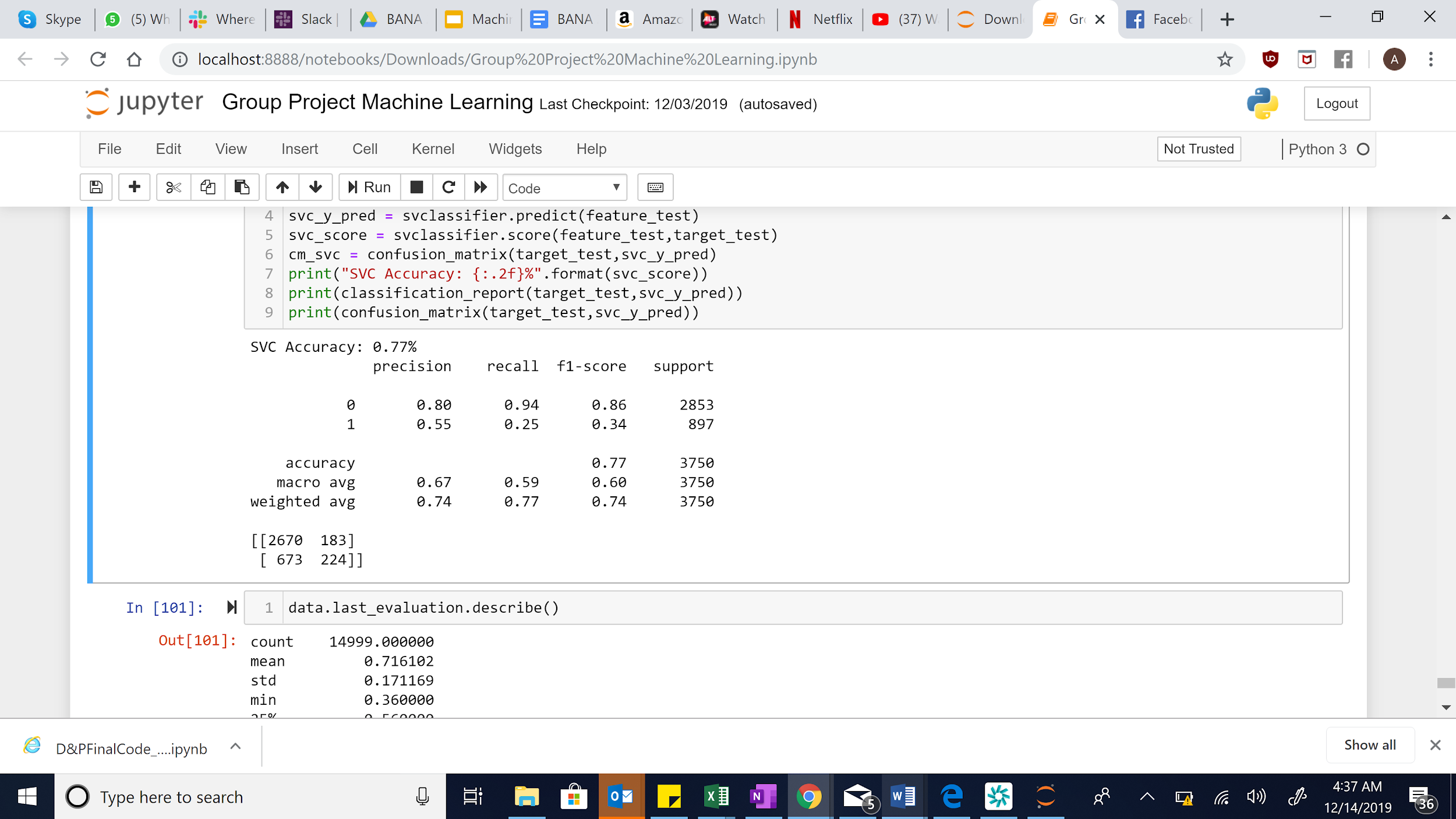
### **K-Nearest Neighbor**

We ran K nearest neighbours classifier with no of neighbours= 2 on the training data(75% records) and fitted on the test data to get a model accuracy of 94% and area under ROC-AUC curve of 92% (just a % more than the decision tree classifier that we are finalizing). Below is the report for reference



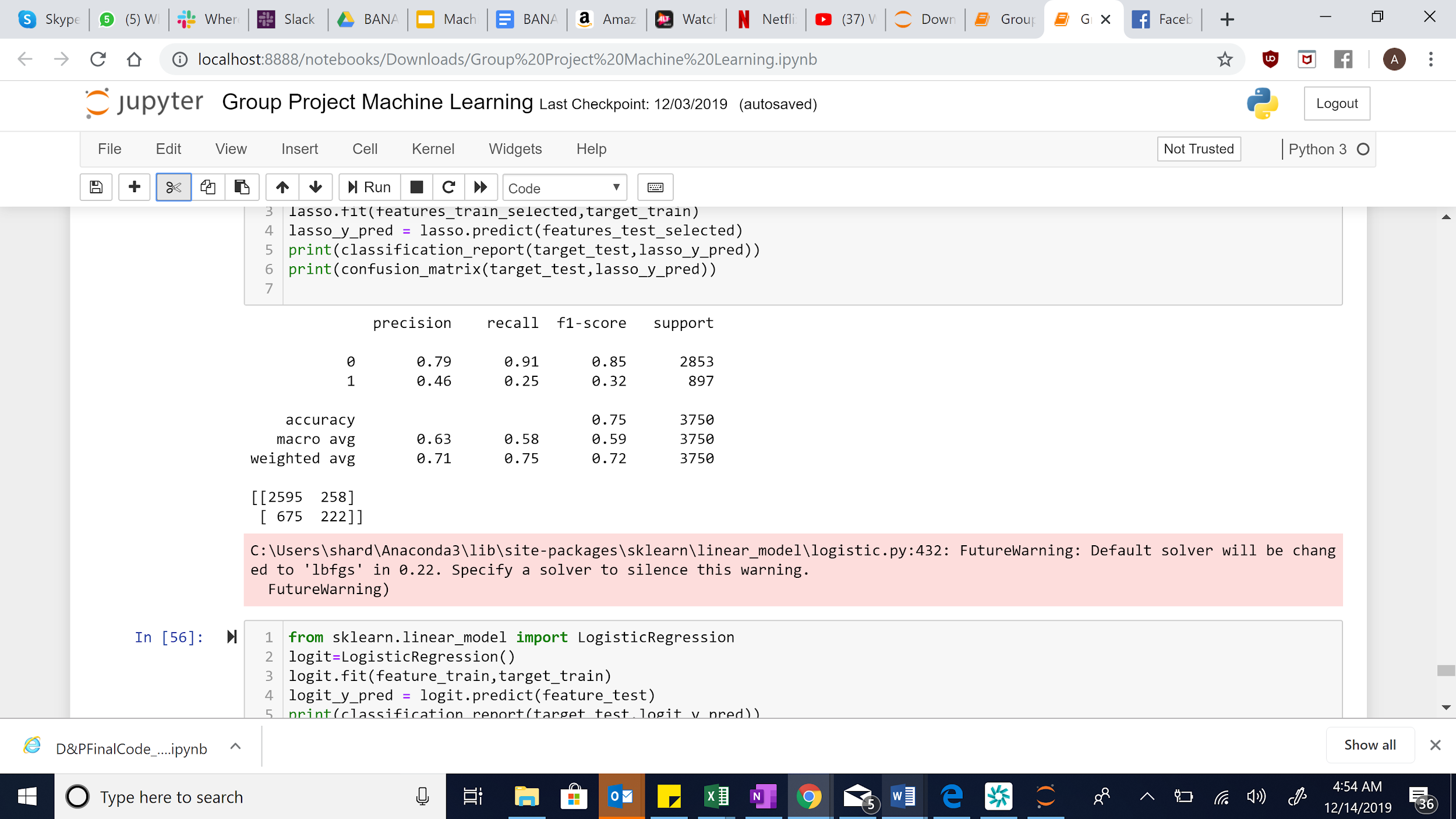
### **Support Vector Machine**

We ran SVC classifier with kernel as linear and got an accuracy of only 77%. Below is the classification report for reference.



### **Lasso Regression**

We also applied a L1 penalty on the logistics regression model with a lambda value of 0.1 obtained by doing cross validation on lasso regression model using lambda value from 0.5 to4.5 and the model with best score corresponding to lambda. We only got an accuracy of 75% from this model. Below is the classification report for reference



### **Figure 1 - HR Analytics Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Attribute Type** | **Description** |
| satisfaction\_level | Numerical  (Float) | Manifests the employee’s satisfaction rating, ranging from 0-1 |
| last\_evaluation | Numerical  (Float) | Manifests the employee’s most recent evaluation rating, ranging from 0-1 |
| number\_project | Numerical  (Integer) | Indicates the number of projects the employee conducted |
| average\_montly\_hours | Numerical  (Integer) | Denotes the average monthly hours the employee worked |
| time\_spend\_company | Numerical  (Integer) | Specifies the number of years the employee spends at the company |
| Work accident | Categorical | Demonstrates whether the employee experienced a work accident at this particular company. 1 = a work accident occurred and 0 = no work accident occurred |
| left | Categorical | The target feature for the model, describing whether an employee has left the company with 1 = employee has left the company and 0 = employee has not left the company |
| promotion\_last\_5years | Categorical | Indicates whether an employee has received a promotion in the last 5 years. 1 = yes and 0 = no |
| sales | Categorical | Describes the employee’s respective department in the company such as sales, technical, product management, IT, etc. |
| salary | Categorical | Illustrates the employee’s salary levels: low, medium, or high |

### **Figure 2 - Description of the Confusion Matrix for the Dataset**

Example:

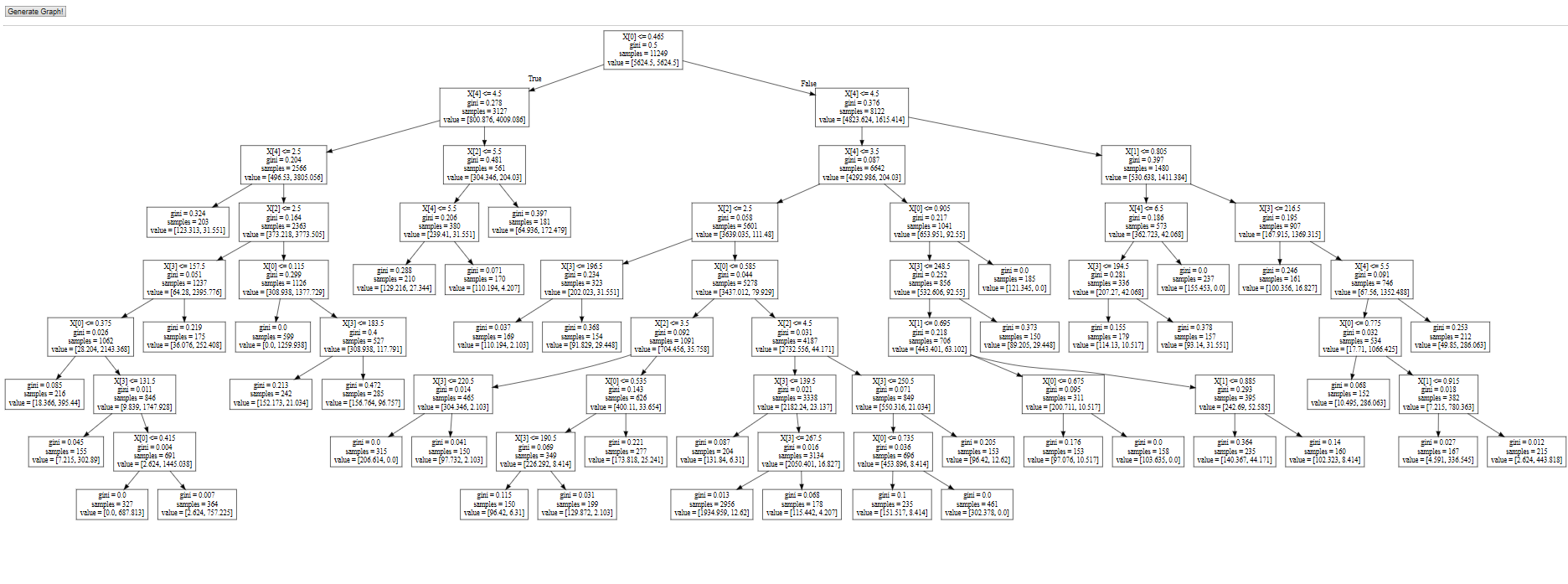
a b

2132 (TN) 149 (FP) | a = 0

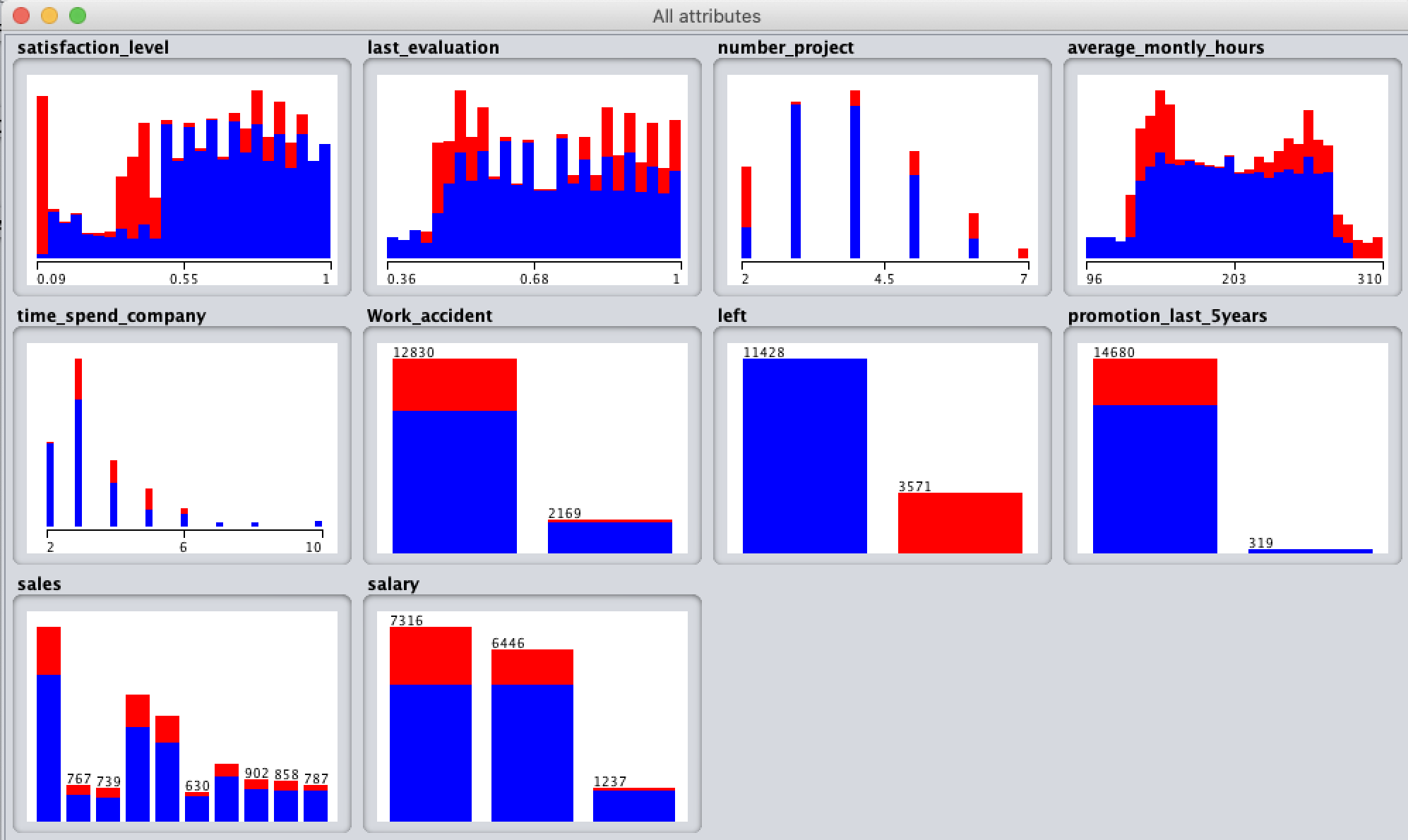
460 (FN) 259 (TP) | b = 1

|  |  |
| --- | --- |
| **Confusion Matrix** | **Description** |
| True Positive | Model predicts employee is likely to leave the company and the employee left the company in reality. |
| True Negative | Model predicts employee is not likely to leave the company and the employee did not leave the company in reality. |
| False Positive | Model predicts employee is likely to leave the company, but the employee **did not** leave the company in reality. |
| False Negative | Model predicts employee is not likely to leave the company, but the employee **did** leave the company in reality. |

### **Figure 3 - Decision Tree (Depth 7)**



### **Figure 4 - visuals of all features with the target class (left) in Weka**



**END OF DOCUMENT**