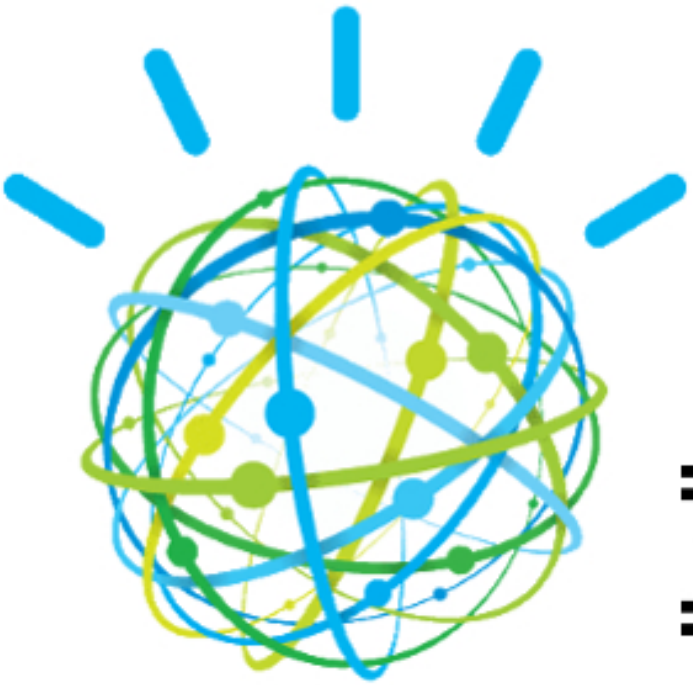


IBM HR Analytics Employee Attrition

(Improved Analysis with Oversampling, PCA & Kernel SVM)

1. *Introduction*

This case study is a continuous study based on my Case Study 1. This time, I applied new techniques such "PCA", "SMOTE oversampling" and "Kernel SVM" to engineer new features and to improve predicting performance for employee attrition. Meanwhile, I also removed three most important features ("OverTime", "MonthlyIncome", "StockOptionLevel") from the previous study to form a new data set for this analysis.

The employee attrition analysis for Human Resources Analytics has caught more and more attention in the business world, especially for big company, how to use analytic methods to predict whether employees will leave or not can help the company improve the HR management and save the cost on it. Therefore, for the human resources managers, it is crucial to have a better idea of what kind of employees will tend to leave and what kind of features will influence them to leave. Most commonly, companies already engage in using analytic methods to analyzing the employee attrition such as employee surveys. But the result generated by the survey may not tell the truth, employees could lie to the company and hide their intention of leaving. Other methods will rely on the statistics results such as attrition rate to predict how many employees will leave. But this is not enough, here coming the play of the most prosperous trend of the big data era, which is the machine learning technique.

This improved study is aiming to:

1. Whether to use machine learning algorithm along with oversampling technique and PCA for extracting new features could yield a better prediction
2. Find the new underlying features and how these features will influence employees' attrition

2. *Data Description*

This study will use the IBM HR Analytics Employee Attrition & Performance data set to apply the machine learning technique to dive into and find the insights of the data. This data set is from Kaggle and it is a fictional data set created by some IBM data scientists with the purpose of mining the insights that lead to employee attrition. This data set consists of 1470 instances, which including 1 target variable "Attrition" ("Yes" or "No") and 31 predicting attributes. There is no missing value in the data set. The target attribute contains 1233 "No" cases and 237 "Yes" cases, which might lead to a data imbalance problem during the data mining process. Here is a table of the data description:

Feature	Data Type	Description
Age	Continuous	Employee's age, Years
Attrition	Categorical	Target variable, 2 levels, "Yes" or "No"
BusinessTravel	Categorical	How often does employee travel, 3 levels: Non, Frequently, Barely
DailyRate	Continuous	Employee pay rate per day
Department	Categorical	Employee's department, 3 levels, "HR", "R&D", "Sales"
DistanceFromHome	Continuous	Employee's distance from home to work
Education	Ordinal	Employee's educational level, 5 levels
EducationField	Categorical	Employee's educational field of study, 6 levels
EnvironmentSatisfaction	Ordinal	Employee's satisfaction with work environment, rating, 1-4
Gender	Categorical	2 levels
HourlyRate	Continuous	Employee's hourly rate
JobInvolvement	Ordinal	Employee's job involvement level, rating, 1-4
JobLevel	Ordinal	Employee's job level of experience, 5 levels
JobRole	Categorical	Employee's specific job role, 9 levels
JobSatisfaction	Ordinal	Employee's job satisfaction level, rating, 1-4
MaritalStatus	Categorical	Employee's marital status, 3 levels
MonthlyRate	Continuous	Employee's monthly pay rate
NumCompaniesWorked	Continuous	Number of companies employee had worked for previously
PercentSalaryHike	Continuous	Percent employee's salary has been raised
PerformanceRating	Ordinal	Employee's performance rated by manager, rating, 1-4
RelationshipSatisfaction	Ordinal	Employee's satisfaction with their relationship, rating, 1-4
TotalWorkingYears	Continuous	Employee's total working years
TrainingTimesLastYear	Continuous	Number of training times employee received last year
WorkLifeBalance	Ordinal	Employee's rating for their work-life balance, 1-4
YearsAtCompany	Continuous	Employee's working year in the company
YearsInCurrentRole	Continuous	Employee's working year in the current role
YearsSinceLastPromotion	Continuous	Years since last promotion
YearsWithCurrManager	Continuous	Years with current manager

3. ***Data Cleaning***

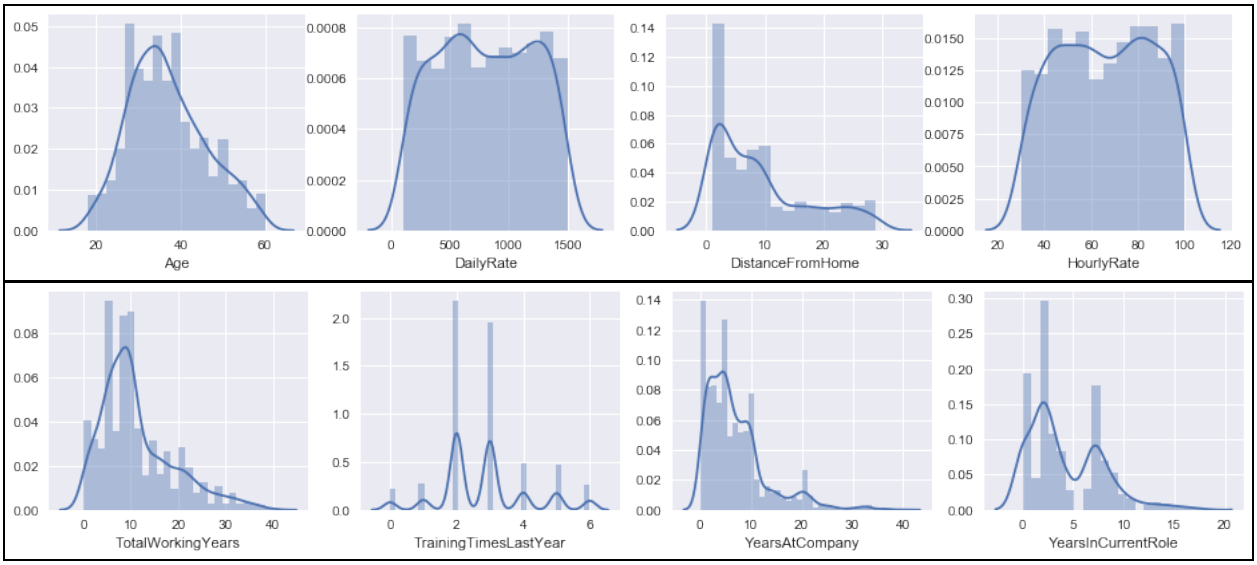
Before starting the data mining process, I conducted some data cleaning for the data set. Among all the 32 features in the data set, there are 4 features that are not useful, they are they are “EmployeeCount” (count the number of employees), “EmployeeNumber” (employee ID number), “Over18” (all the employee are over 18), and “StandardHours” (all the values are 80). Therefore, I removed these 4 features and kept 28 features including 27 predictors and 1 target variable.

During the data mining process, I created dummy variables for the categorical features in order to apply the machine learning algorithms more smoothly on the data set. After transforming the dummy variables, I got 41 features in total. Then I conducted the PCA to extracted 28 most important features.

4. ***Data Analysis***

There are 3 types of data in this data set, such as continuous, categorical and ordinal features. For the purpose of fully exploring all kinds of features and find the underlying relationship within features, I performed different data visualization techniques for each kind of feature and correlation analysis to reveal them.

Figure 1 Histograms for Continuous Features



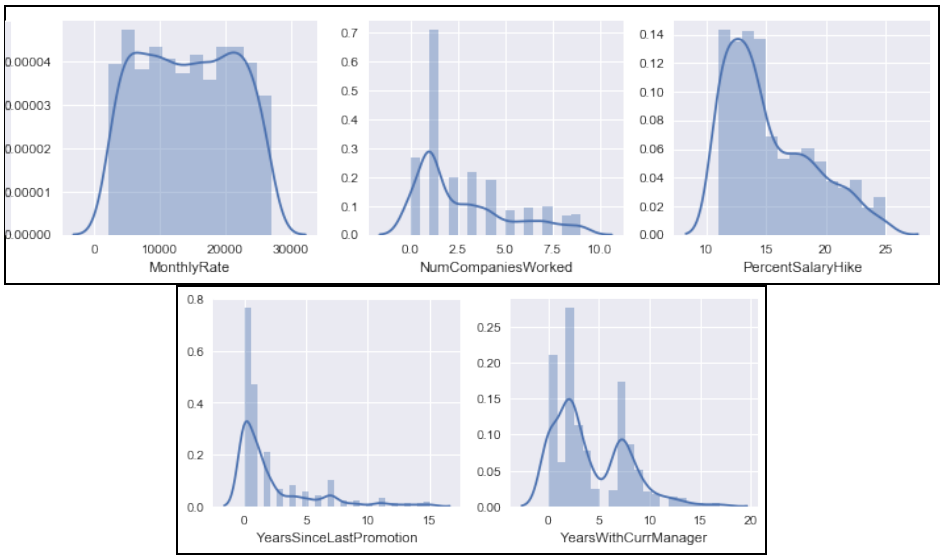


Figure 2 Boxplots for Continuous Features

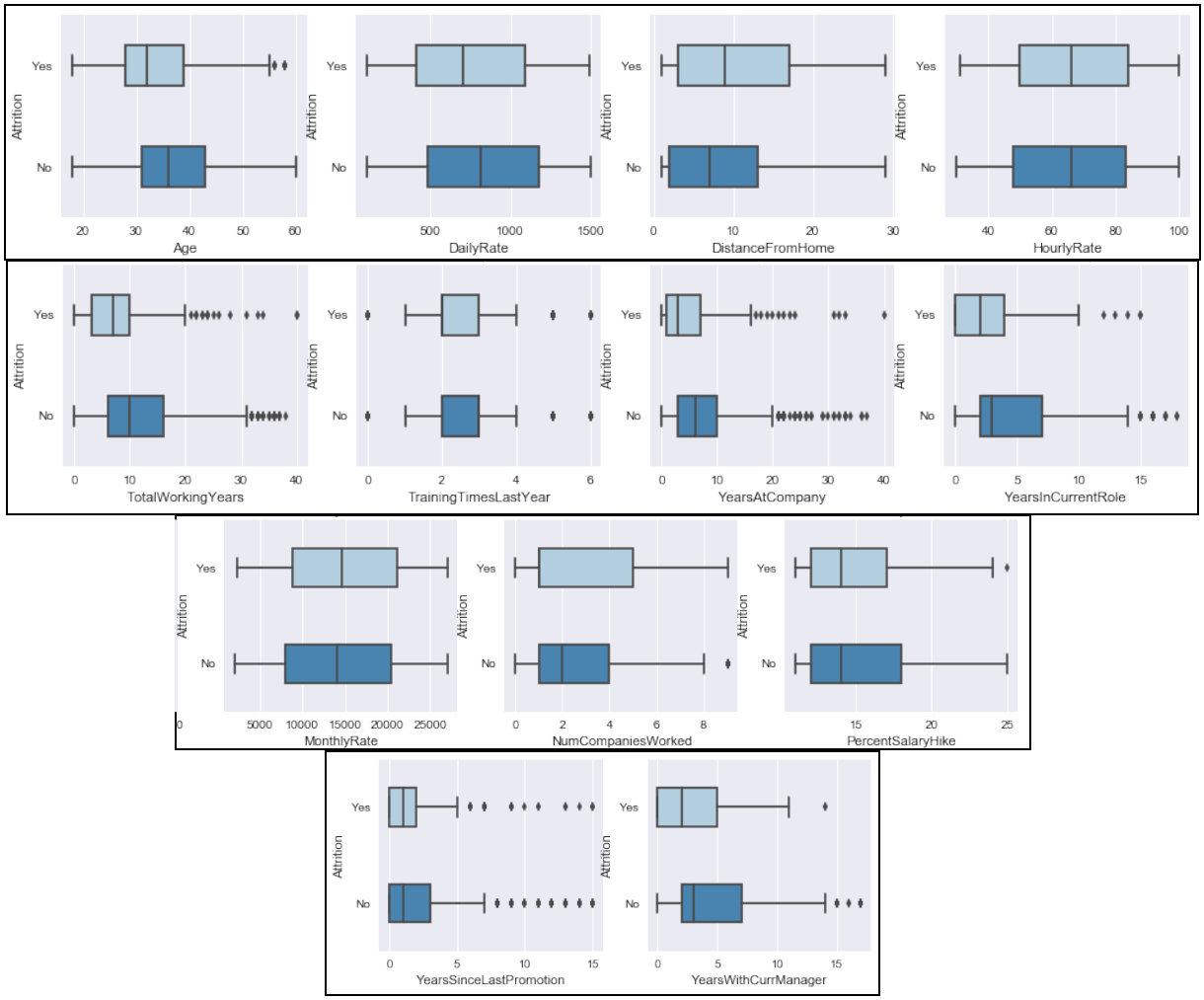
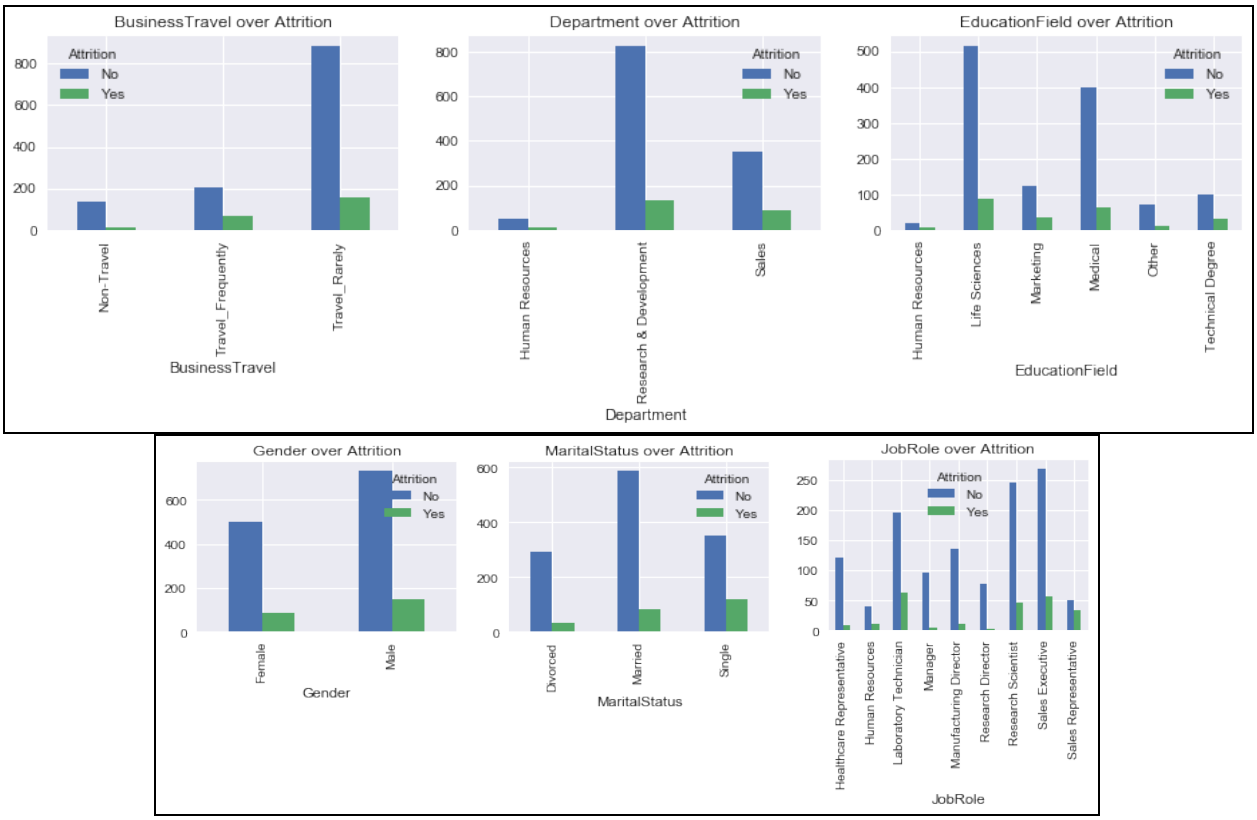


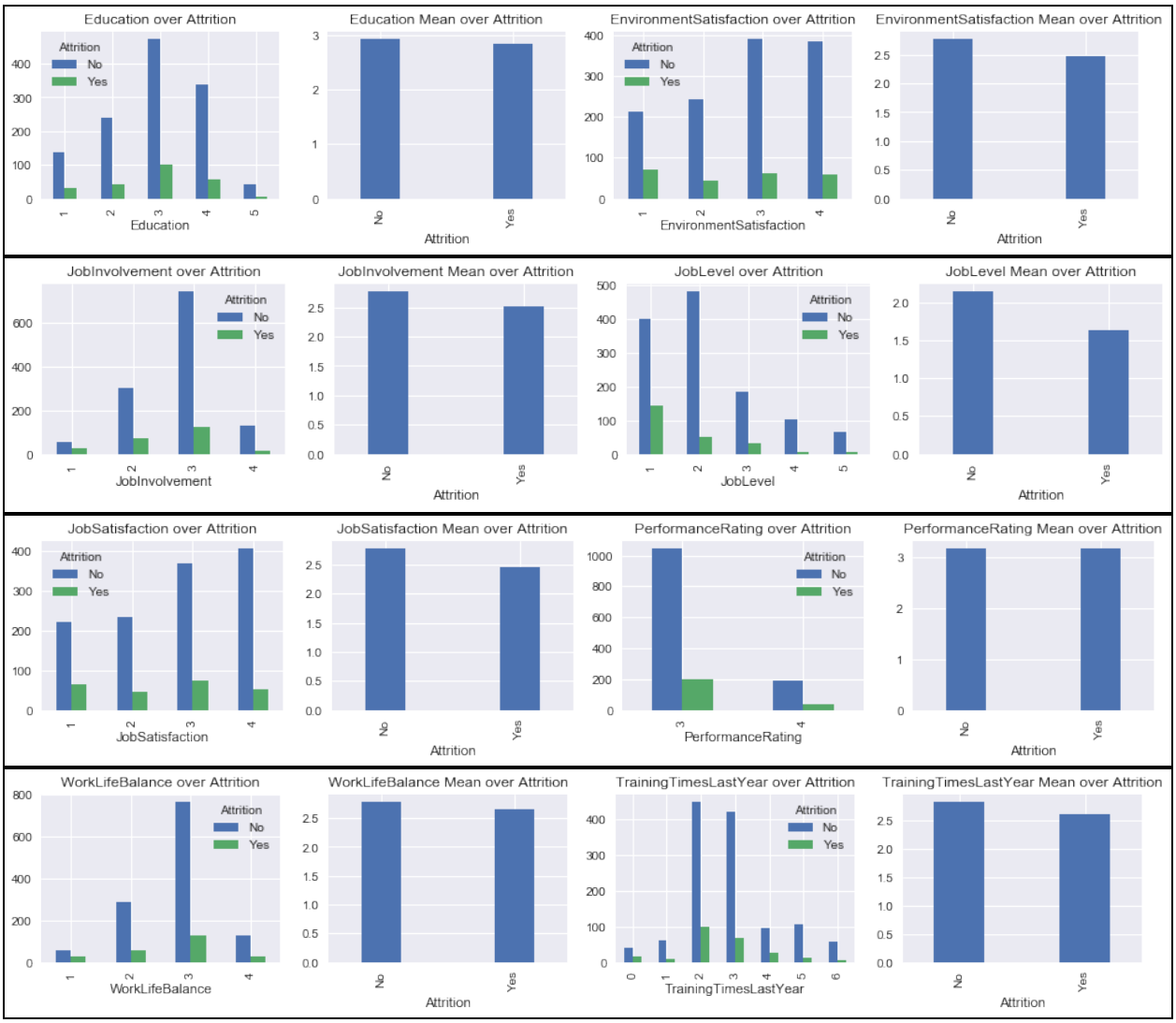
Figure 2 shows the underlying differences between those who leave the company and who stay at the company. We can see from the plots that the average age of the employees who leave is younger than the employees who choose to stay, which means that younger people are more likely to change jobs. And employees who leave mostly has a lower monthly income as well daily rate than employees who stay. For some other features such as “TotalWorkingYears”, “YearsAtCompany”, “YearsCurrentRole” and “YearsWithCurrManager”, they all shows that employees who stay and work for the company for a shorter time are more likely to leave the company, while employees who work longer for the company are less likely to leave.

Figure 3 Cross-tabulation Bar Plots for Categorical Features



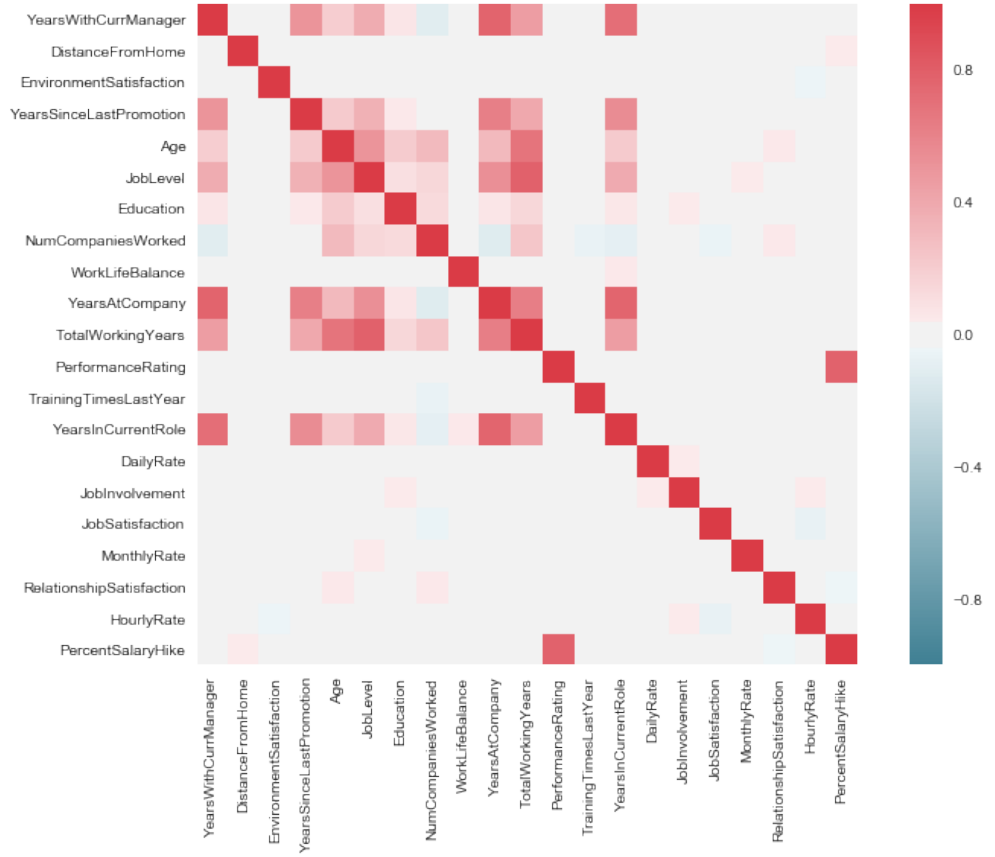
For the categorical features, I applied the cross-tabulation frequency bar plots to show the differences between the employees who leave and stay within each category. From the plots, we can see that for the employees who travel a lot have a higher attrition rate, and for the employees who have overtime also have a higher attrition rate. Besides this, we can find that for different job roles, sales representative has a very high attrition rate. The “JobRole over Attrition” bar plot shows that for the sales role, there are about 40% of the sales representatives leave the company.

Figure 4 Cross-tabulation Bar Plots for Ordinal Features



For the ordinal features related to rating or satisfaction measure, I first applied the cross-tabulation in order to find differences and interesting patterns among them. But it doesn't show many interesting insights since the target feature is too imbalanced. Then I used the bar plot to compare every ordinal feature's mean between employees who leave and not leave, it gave me some useful information. From the plot, it is very obvious to see that employees who are more likely to leave company have lower environment satisfaction and job satisfaction, lower job level and stock option level. But there is no significant difference in the education level.

Figure 5 Correlation Plot for Continuous Features



After taking out three important features from the data set, the correlation plot changed a little bit compared to the last analysis. From the above correlation plot, I found that there were several very strong positive correlated relationships, they were “JobLevel” along with “TotalWorkingYears”, “YearAtCompany” along with “YearsWithCurrentManager”, and “PercentSalaryHike” along with “PerformanceRating”. Based on the correlation analysis, I conducted the PCA on the data set. I found there were 28 principal components that could explain overall 95% of the variance from the data set.

Figure 6 Barplot for Principal Component's Explainded Varinance Ratio

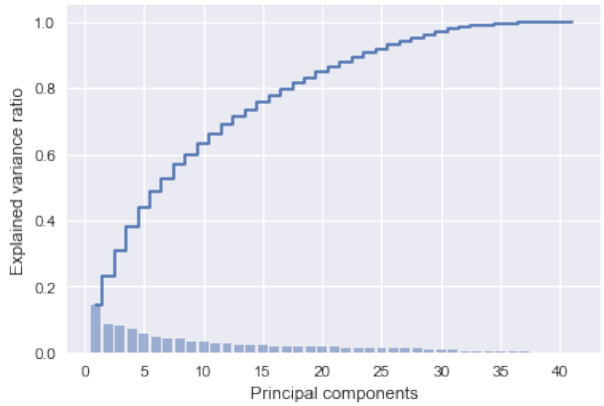


Figure 7 Heatmap for PCA

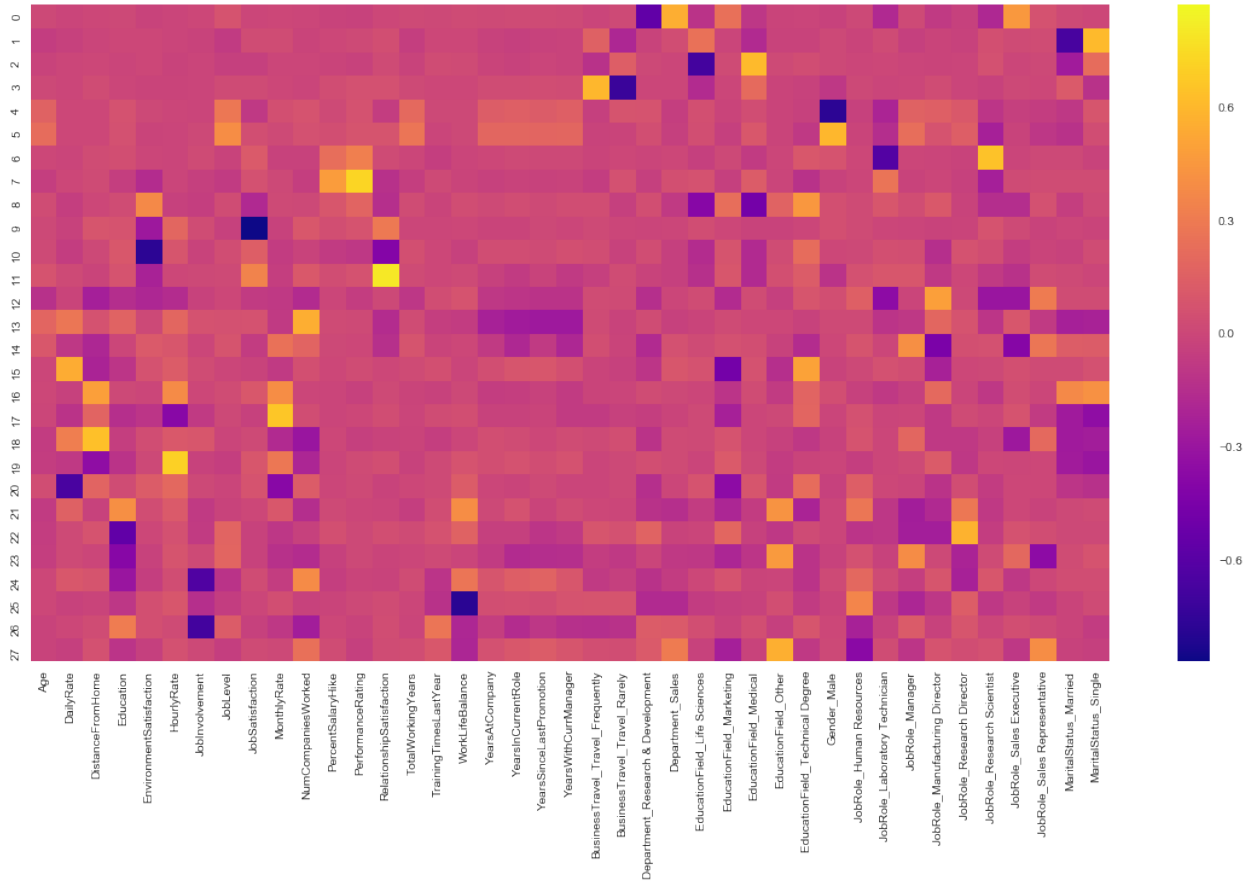
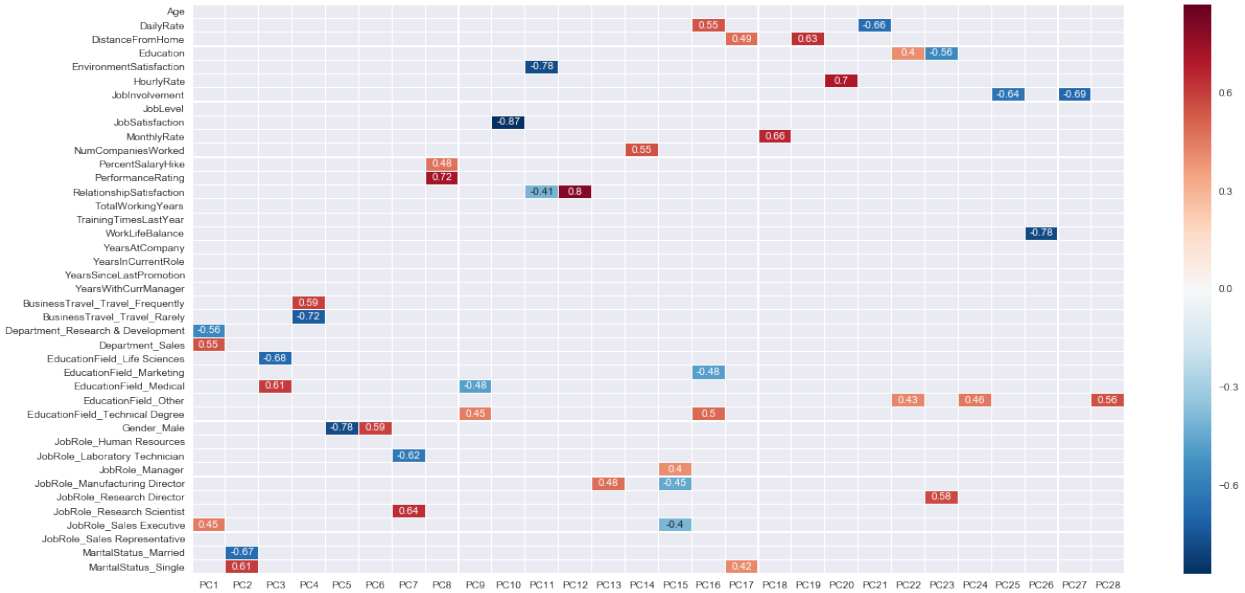


Figure 8 Heatmap for Common Factor Analysis



From the above two plots, we could see the 28 new features after applying the PCA and common factor analysis.

5. ***Experimental Results***

For the experimental analysis, I separated the data set into training and testing set, 80% for the training set and 20% for the testing set. I utilized the stratified 10-fold cross-validation on training set to optimize the model and testing set to compare different classifiers. But there is one difference for this times' analysis. I oversampled my training data during the cross-validation process to optimize the model for solving the data imbalance problem.

For the classifiers, I applied Decision Tree Classifier, Kernel Support Vector Machine, Random Forest Classifier and Stacking Ensemble Algorithm.

For the key performance metrics, I kept using "Recall" and "Accuracy" scores. The reason why to use "Recall" as the main criteria is it is a measure of the true positive class over the actual positive class. A higher "Recall" score will mean that my model could predict more actual employee attrition, which is the main purpose of the analysis.

Decision Tree Classifier:
First, I tried the Decision Tree Classifier with grid search using the stratified 10-fold cross validation to get an experimental result for the classifier. The result was not good; the recall score was only 26.9% with a very complex tree model with 9 maximum depth. From this result, I found that the Decision Tree Classifier suffer a huge over-fitting problem. To overcome this problem, I tuned the Decision Tree Classifier by finding the best parameters before over-fitting occurs using 10-fold cross-validation with oversampling for the training set. After tuning, I got the best parameters with 3 maximum depth, 1 minimum samples leaf, 2 minimum samples split and using "Entropy" as the criterion. I fit the model with these set of parameters on the oversampled training set and got the prediction results as below:

Table 1 Recall & Accuracy Scores for Decision Tree Classifier

DT Cross-validation		DT Test		Recall 95% CI
Recall	Accuracy	Recall	Accuracy	
64.2%	54.3%	51%	56%	(54.4%, 74.1%)

Figure 9 Decision Tree Plot

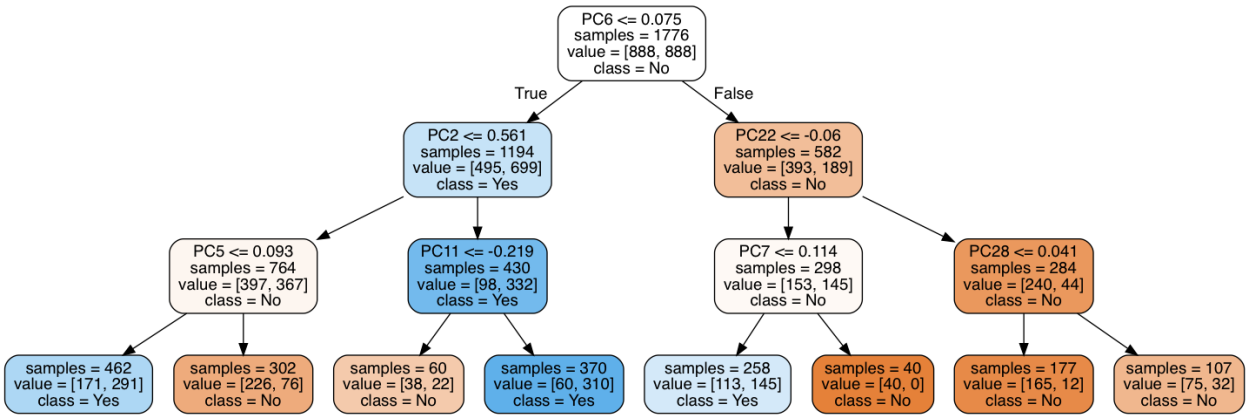
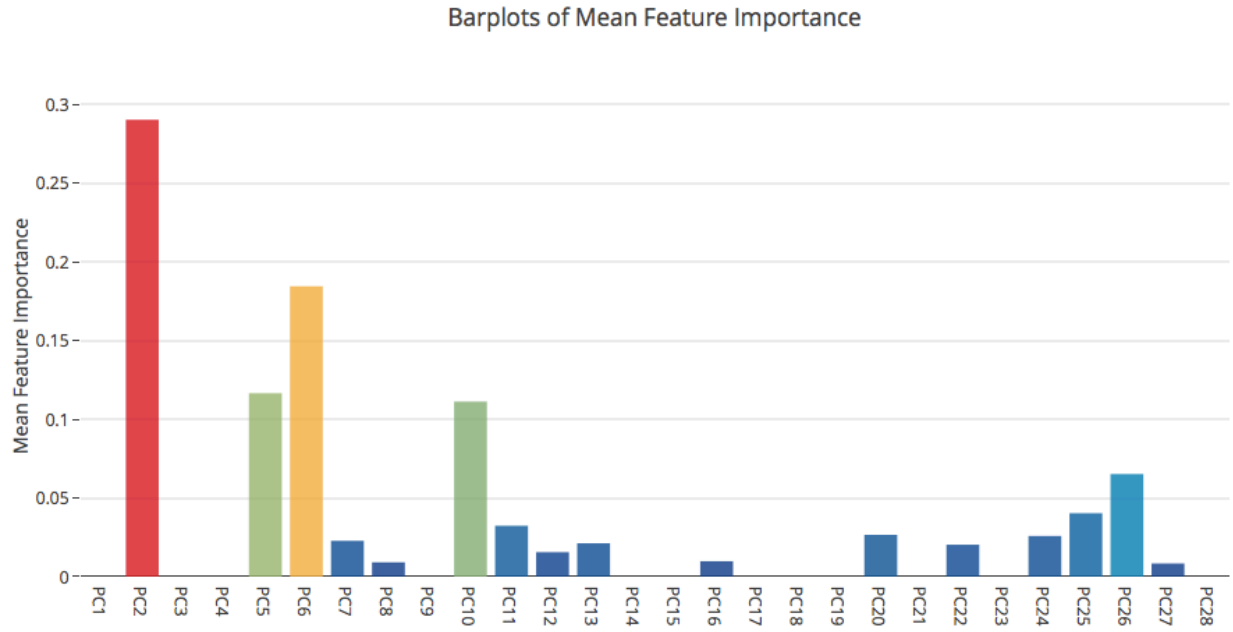


Figure 10 Barplot of Mean Feature Importance



The above plot shows that the mean most important features during the cross-validation process of the Decision Tree Classifier are “PC2 (Marital Status with single)”, “PC6 (Gender Male)”, “PC5 (Gender Female)”, and “PC10 (Low Job Satisfaction)”.

Kernel Support Vector Machine:

For the Kernel SVM Classifier, I tried SVM with “linear”, “polynomial”, “rbf”, and “sigmoid” kernels. As the result, “linear” and “rbf” kernels are much better than the other two kernels. But the difference between the “linear” and “rbf” kernels are pretty small. Therefore, I

compared the “linear” and “rbf” kernels with different “C” value using the cross-validation along with oversampling during the process.

Figure 11 SVM with "Linear" kernel 10-fold Cross-validation Recall over C=[0,1] Plot

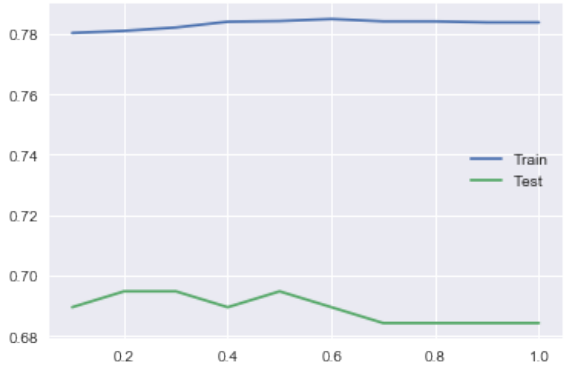
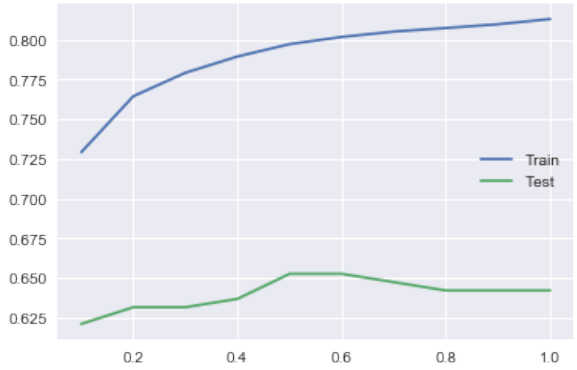


Figure 12 SVM with "RBF" kernel 10-fold Cross-validation Recall over C=[0,1] Plot



From the above two plots, It’s obvious to see that the “linear” kernel have a higher “Recall” score than the “RBF” kernel in general. The best “C” value for the “Linear” kernel is 0.2. To validate whether the difference is significant or not, I conducted a “Paired T-Test” for the two 10-fold cross-validation’s results. The “P-value” of the test is 0.069 that is bigger than 0.5, which means that the there is a significant difference between the two 10-fold cv’s results. Therefore, it can be proved that the “Linear” kernel with “C=0.2” is the best set of parameters for building the model. Finally, the prediction result I got is as below:

Table 2 Recall & Accuracy Scores for “Linear” Kernel SVM (C=0.2)

SVM “Linear” Kernel (C=0.2) Cross-validation		SVM “Linear” Kernel (C=0.2) Test		Recall 95% CI
Recall	Accuracy	Recall	Accuracy	
69.5%	70.6%	62%	71%	(58.7%, 80.3%)

After trying the Decision Tree and SVM classifiers, the better was the SVM with “Linear” kernel classifier. It yielded a “Recall” score for test set of 62% and an accuracy score of 71% after tuning the parameters. Comparing the results from this analysis to the previous analysis for the individual classifiers. The predicting performance with oversampling during the cross-validation process did improve a lot especially for the “Recall” score. The “Accuracy” decreased a little, but it was still acceptable.

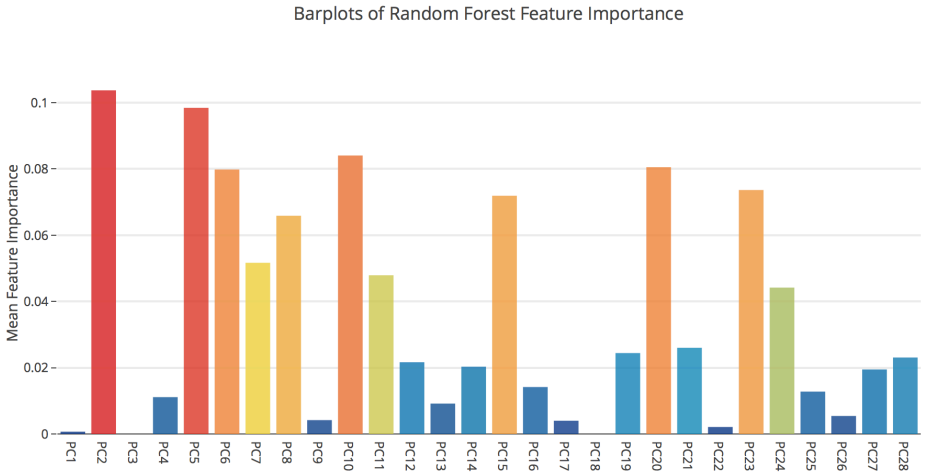
After utilized the individual classifiers, I started to apply the Ensemble Classifiers to see if they could perform better. I tried two ensemble methods, “Random Forest” bagging algorithm and “Stacking” algorithm using Decision Tree and “Linear” kernel SVM.

Random Forest Bagging Algorithm:
Since the Random Forest is based on the Decision Tree Classifier, based on the previous analysis, I have already known there would be an over-fitting problem if I don't tune the parameters. Therefore, I kept using the tuning method as before to find the best parameters before over-fitting problem occurs. The best set of parameters I got were using the 10 as the number of trees in the forest, maximum depth was 4 and “Entropy” as the criterion. Below is the final predicting results for the Random Forest classifier:

Table 3 Recall & Accuracy Scores for Random Forest (n_trees = 10, max_depth = 4)

Random Forest Cross-validation		Random Forest Test		Recall 95% CI
Recall	Accuracy	Recall	Accuracy	
56.3%	67.6%	49%	71%	(47.3%, 65.4%)

Figure 13 Barplot of Random Forest's Feature Importance



From the above barplot, the number of important features for the Random Forest classifier increased a little comparing to the Decision Tree classifier’s. Here we could see that there were more important features besides “PC2”, “PC5”, “PC6” and “PC10”.

Stacked DT, Kernel SVM Classifier:
For the stacking classifier, I used the two individual classifiers used before and used "soft" voting as the criterion. The results I got are showed as below:

Table 4 Recall & Accuracy Scores for Stacked DT and Kernel SVM Classifiers

Stacking Cross-validation		Stacking Test		Recall 95% CI
Recall	Accuracy	Recall	Accuracy	
66.3%	72.7%	62%	72%	(54.5%, 78.1%)

The stacked classifier yielded a good result almost same as the Kernel SVM classifier, but was better than the Decision Tree and Random Forest classifiers.

6. *Experimental Analysis*

Table 5 Test Scores for Different Classifiers

Classifier	**Test Scores**				**Over Fitting**
	Recall	**Precision**	**Accuracy**	**AUC**	
Decision Tree	51%	18%	56%	60%	Y
Kernel SVM	62%	30%	71%	72%	N
Random Forest	49%	27%	71%	64%	N
Stacking	62%	31%	72%	71%	N

After performing all the experiments above, it can be seen that both the "Kernel SVM" and Stacking Classifier yielded almost the same good results with "Recall" of 62% and "Accuracy" of 71% and 72%. And their "AUC" are also nearly the same as each other with all above 70%.

The over-fitting problem occurred severely in the last analysis only occurred on the Decision Tree Classifier for this analysis after applying the oversampling method on the data set during the cross-validation process to build the model. Besides the oversampling, another reason that contributed in solving the over-fitting problem could due to using PCA to extract 28 new features for prediction. After utilizing the PCA feature extraction, the number of total features decreased from 41 to 28. Combining these two methods, the over-fitting problem was solved since by using less features along with more samples.

In conclusion, I think the best classifier to predict employee attrition for this case should be the "Kernel SVM" since it has less computational cost compared to the "Stacking" (stacked DT and kernel SVM). But I think the Random Forest could also be considered to use if the data set becomes very large. Although the "Kernel SVM" yielded a better prediction results, the

computational cost would increase along with the data set size. Therefore, if the size of the data set is large, we could still consider to use “Random Forest” to make the prediction.

7. Conclusion

After conducting all the above analysis, and get back to the questions from the introduction. For the first question, we can say that for this analysis, after applying the oversampling technique and conducting the feature extraction, we could get a better prediction result for predicting employee attrition. Therefore, we could say that the capability of how accurately one machine learning algorithm can achieve really depends on the data set’s quality such as the amount of instances and the quality of features. And in real world, when dealing with the data analysis related to business problem, we should consider more beyond the accuracy. In this domain, the accuracy isn't the best metrics, in stead, recall is better one.

For the second question, we can say that the new underlying features that will influence whether employees will leave or stay could be: “PC2 (Marital Status with divorce or single)”, “PC6 (Gender Male)”, and “PC10 (Low Job Satisfaction)”. Therefore, employees who are divorced or single male with lower job satisfaction, are more likely to leave the company according to this analysis.