Principal Component Analysis - IBM HR Employee Attrition

A. Goal

Principal Component Analysis (PCA) is used to reduce the number of features needed to be used in a prediction model. In this exercise, we will see the impact of PCA – in terms of accuracy and time needed to train – on models: Logistic Regression, Decision Tree, K Nearest Neighbors, Naive Bayes, and SVM.

B. Data Source

https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

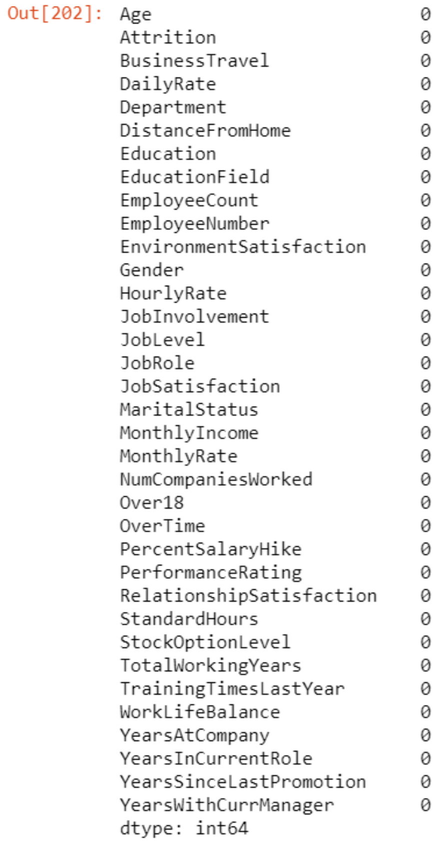
C. Summary

Dataset Description

‘Attrition’ (binary) is the target variable, which we will try to predict. We have 34 input variables – overall including factors such as demographics, travel, education, income, field, and number of years at the current company and role, and with current manager. The features have only 2 datatypes: integers and factors; however, it is important to note that many numerical variables (such as ‘DistanceFromHome’) are ordinal.

Data Cleaning

- We did not have any missing values within the dataset.



```
Out[202]: Age                         0
          Attrition                   0
          BusinessTravel              0
          DailyRate                   0
          Department                  0
          DistanceFromHome            0
          Education                   0
          EducationField              0
          EmployeeCount               0
          EmployeeNumber              0
          EnvironmentSatisfaction     0
          Gender                      0
          HourlyRate                  0
          JobInvolvement              0
          JobLevel                    0
          JobRole                     0
          JobSatisfaction             0
          MaritalStatus               0
          MonthlyIncome               0
          MonthlyRate                 0
          NumCompaniesWorked          0
          Over18                      0
          OverTime                    0
          PercentSalaryHike           0
          PerformanceRating           0
          RelationshipSatisfaction    0
          StandardHours               0
          StockOptionLevel            0
          TotalWorkingYears           0
          TrainingTimesLastYear       0
          WorkLifeBalance             0
          YearsAtCompany              0
          YearsInCurrentRole          0
          YearsSinceLastPromotion     0
          YearsWithCurrManager        0
          dtype: int64
```

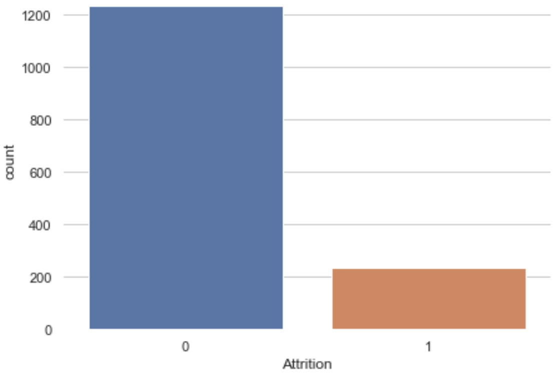
- We mapped the yes/no in the target variable – Attrition – to 1/0.

- We dropped the ‘EmployeeNumber’ as it does not have any impact.

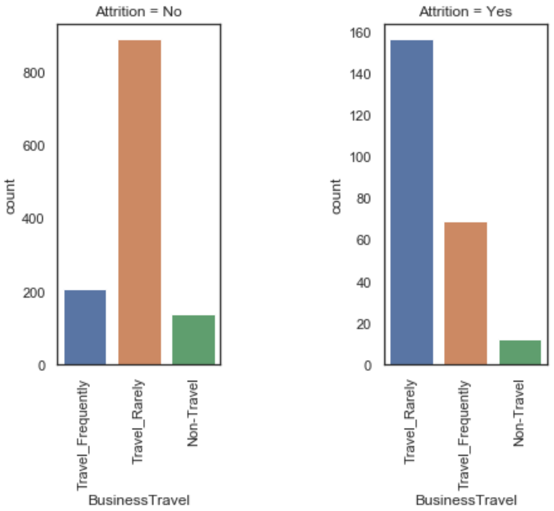
- Further, after looking at the distribution of the numerical variables, we observed that ‘EmployeeCount’ and ‘StandardHours’ are constants; hence, we will drop these columns as well

Exploratory Data Analysis

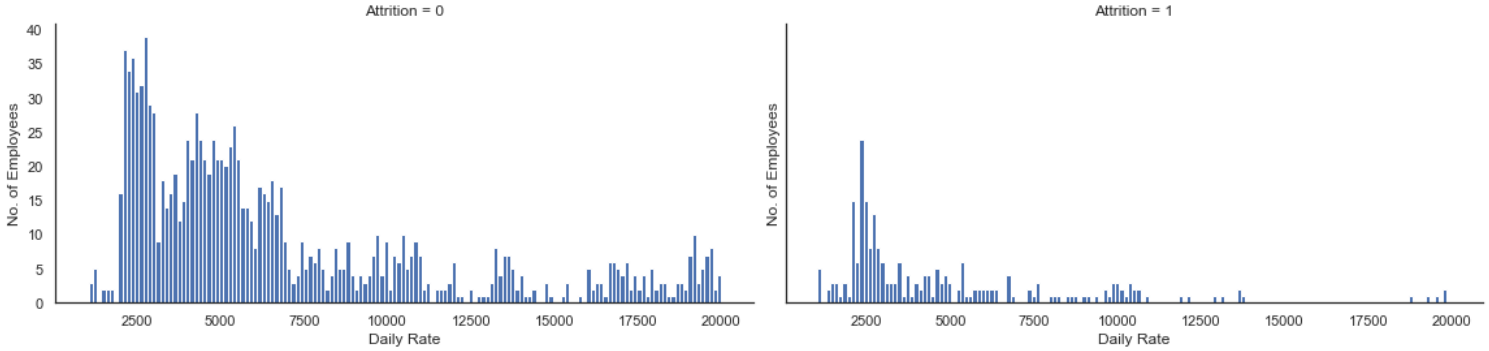
1. We can observe that 'Attrition = 1' is just 1/6th of 'Attrition=0'; only 16% of the entire dataset is '0'. Thus, this is an unbalanced dataset. However, as we are focusing majorly on understanding the impact of PCA, we will not focus on methods needed to deal with unbalanced data (we have already discussed in one of the other repositories (https://github.com/Nickssingh/SMOTE-Unbalanced-Data-Bank Marketing/blob/master/Bank%20Marketing-Call%20Response%20Prediction.Rmd) how SMOTE can be used to deal with imbalance in data)

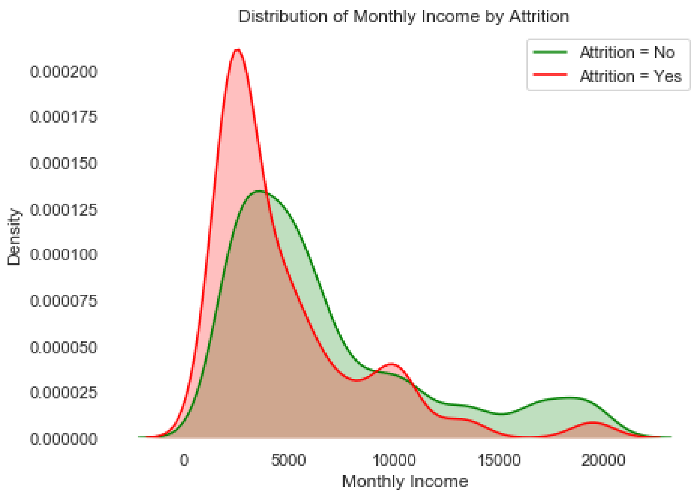


2. It appears that among the employees who discontinued, frequent travel is comparatively a larger portion of the group

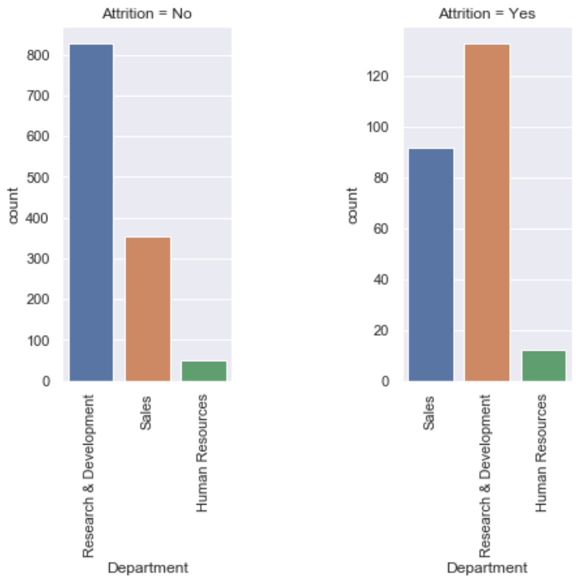


3. We can see that among the employees who discontinued, the higher end of Monthly Income is comparatively sparse, and we visualized this information much more effectively using Kernel Density Estimate (KDE) plots.

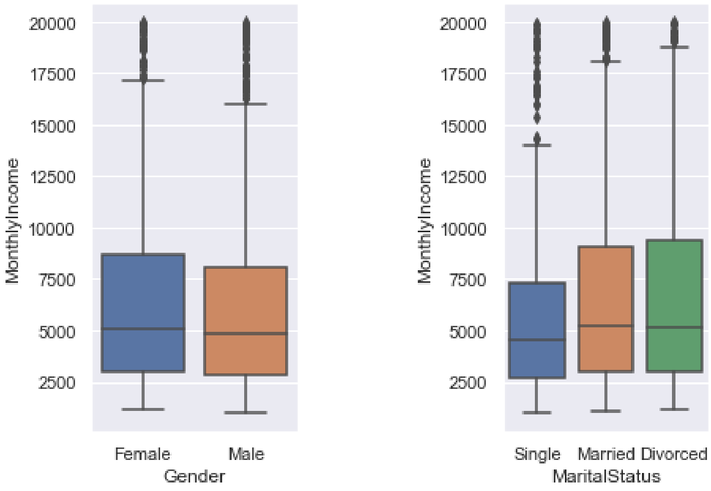




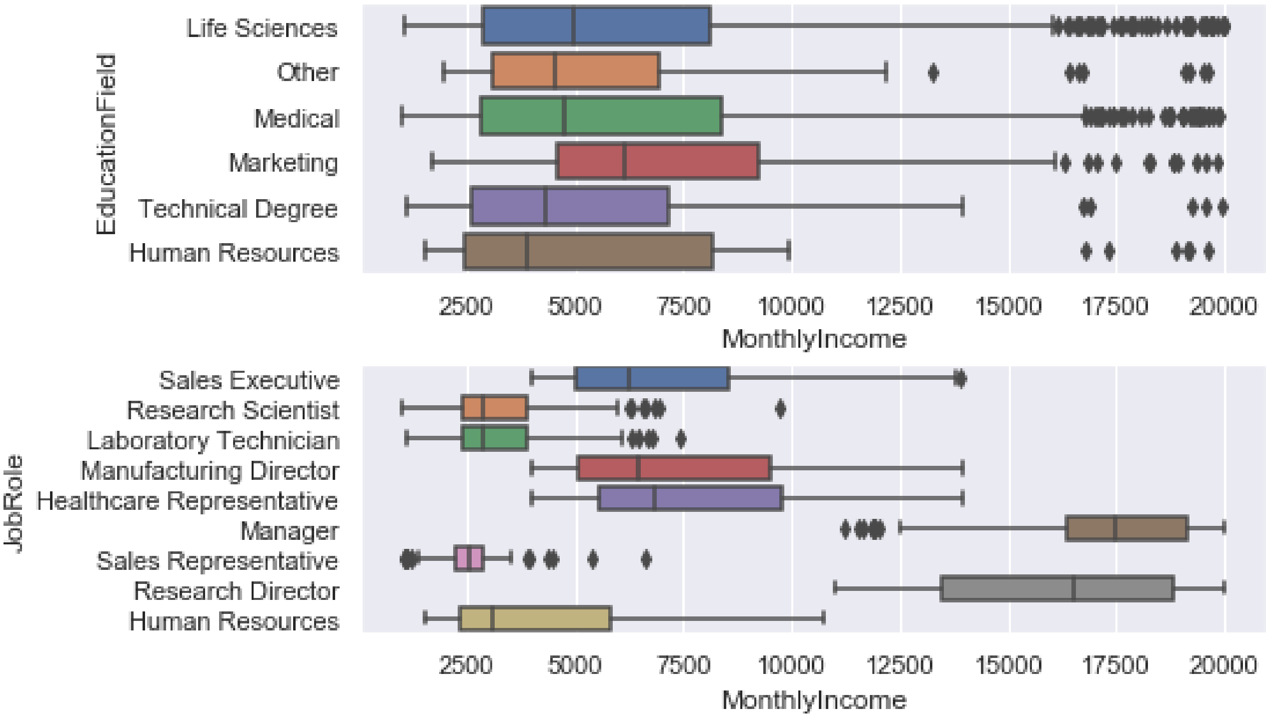
4. The proportion of employees from sales department seems to be higher among those who discontinued



5. We also found that income does not vary widely with gender
6. Single employees have income lower than those of married and divorced ones – probably because single employees are in young and thus in the initial stages of their careers



7. Employees from different education fields seem to have similar income levels; Nonetheless, it can be noted that employees from Life Sciences and Medical background have more number of outliers
8. Research Director and Manager have the highest income levels, and Research Scientist, Laboratory Technician, and Sales Representatives have lowest income levels



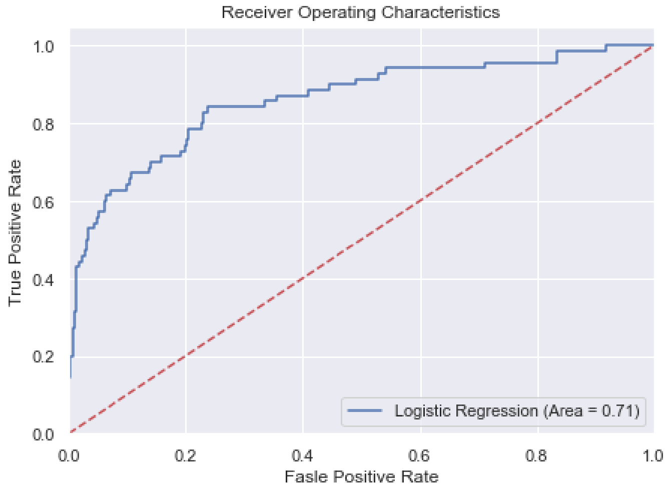
Machine Learning Models with and without Principal Component Analysis

As mentioned in the goal, we developed 5 models. We developed a set of these 5 models without PCA and the other set with PCA. Then, we compared the results in the two sets, using train time and accuracy.

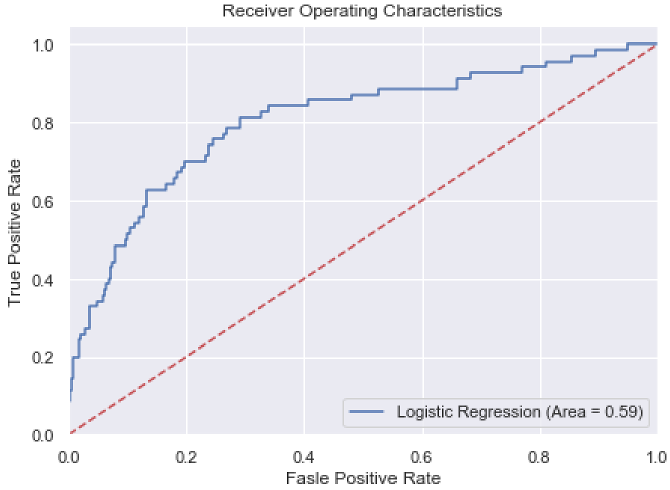
ROC Curves

Logistic Regression

Without PCA

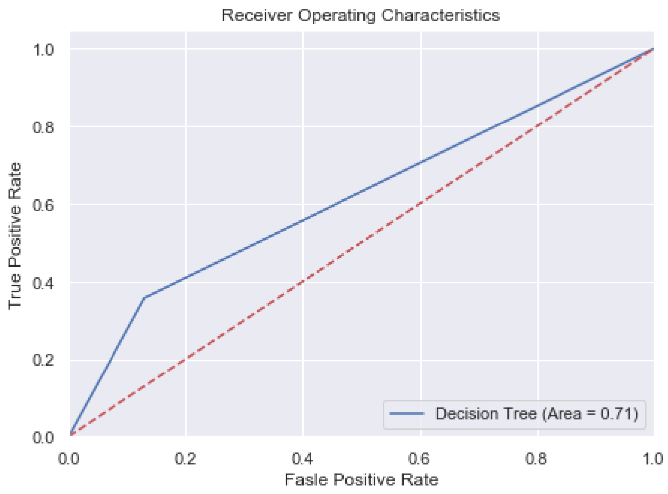


With PCA

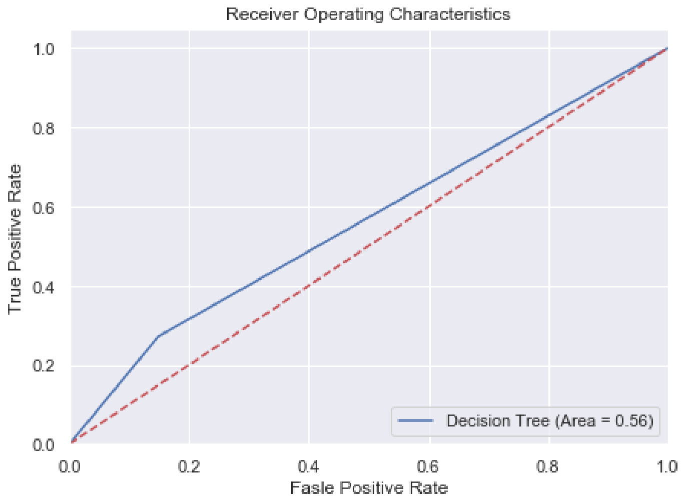


Decision Tree

Without PCA

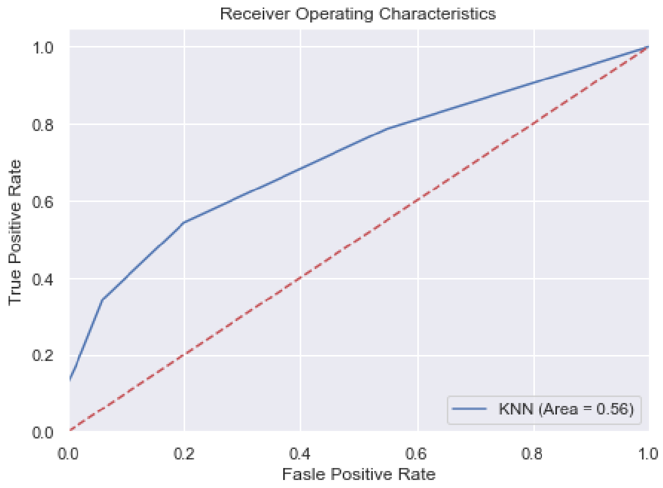


With PCA

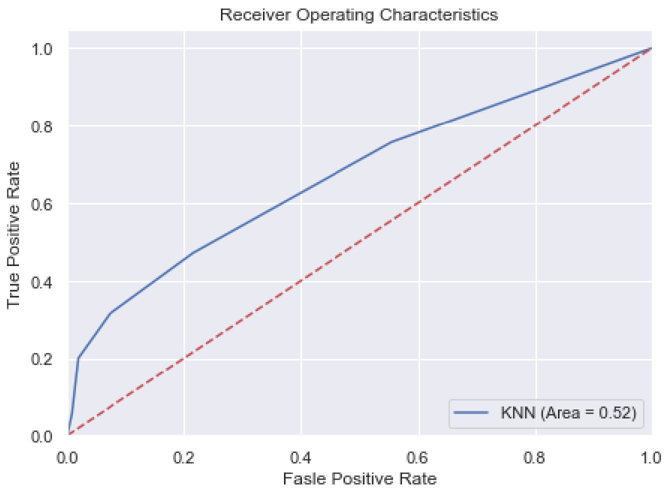


KNN

Without PCA

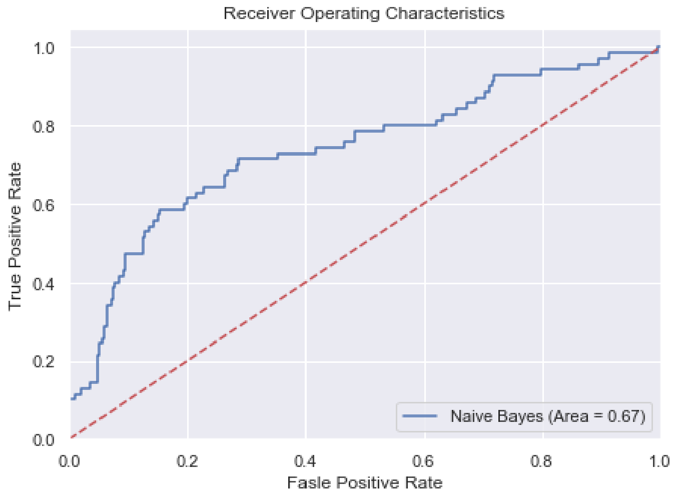


With PC

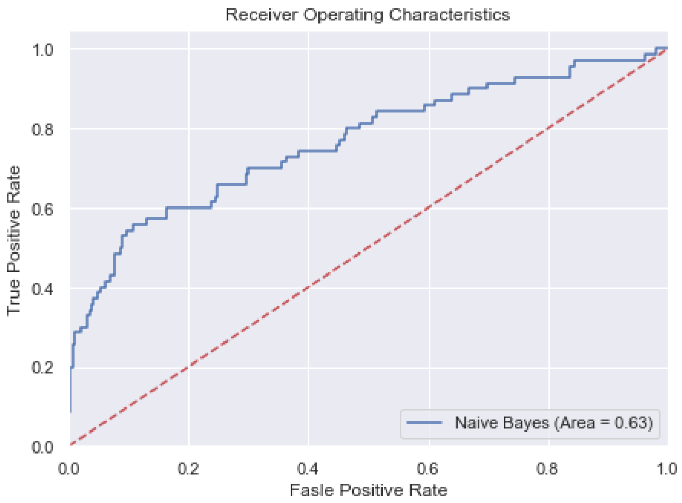


Naive Bayes

Without PCA

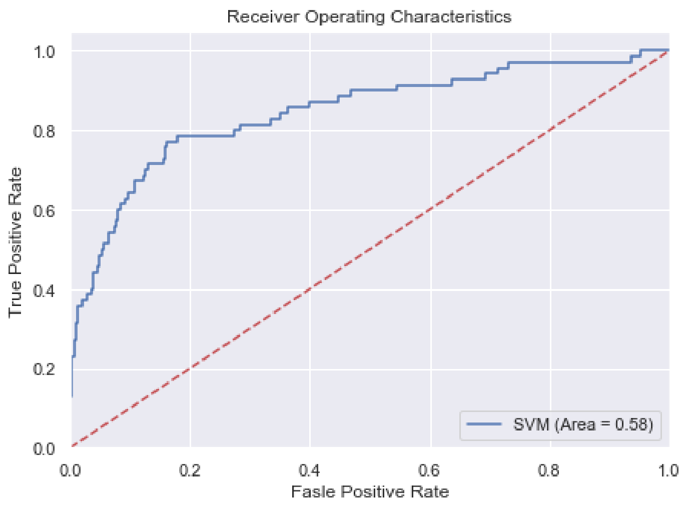


With PCA

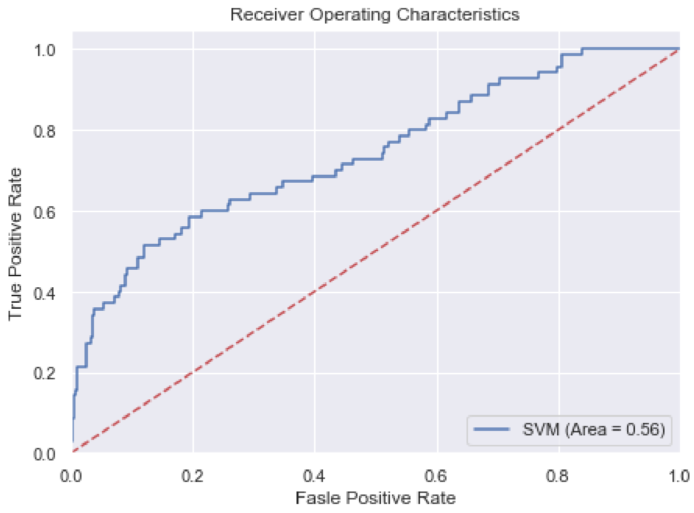


SVM

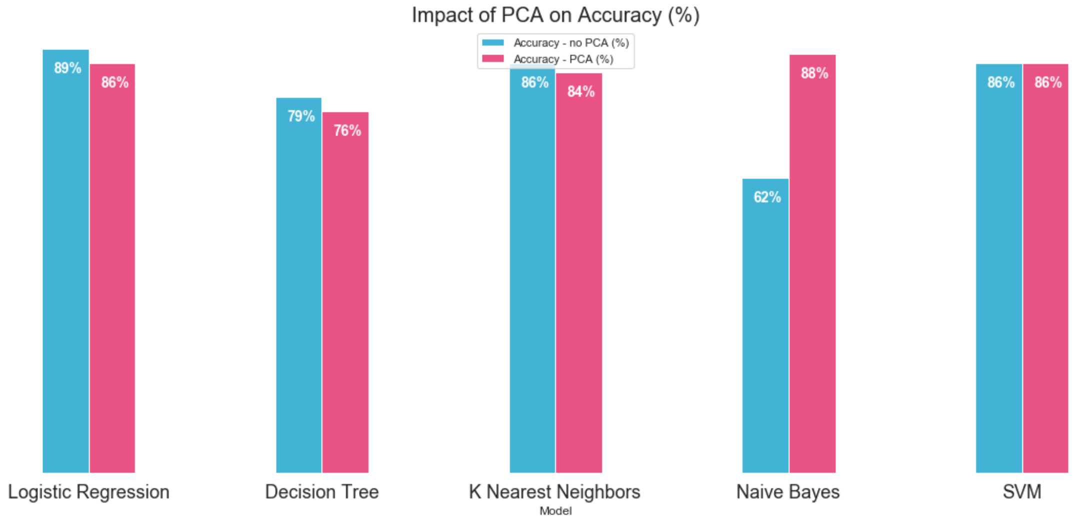
Without PCA



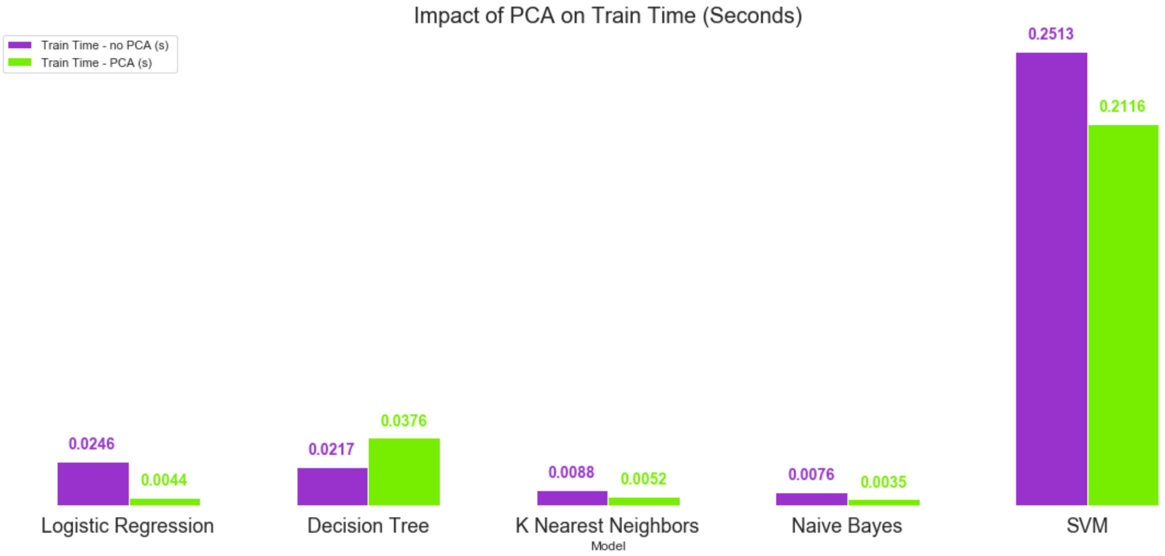
With PCA



Accuracy



Train Time



As shown above, the impact of PCA on Logistic Regression, KNN, and SVM was in alignment without our expectations – decrease in train time with negligible to small compromise on accuracy; the expected change in accuracy was not huge because the features, after reducing the number of variables using PCA, were good enough to explain 95% of the variation. However, the impact on Naive Bayes and Decision Tree was interesting.

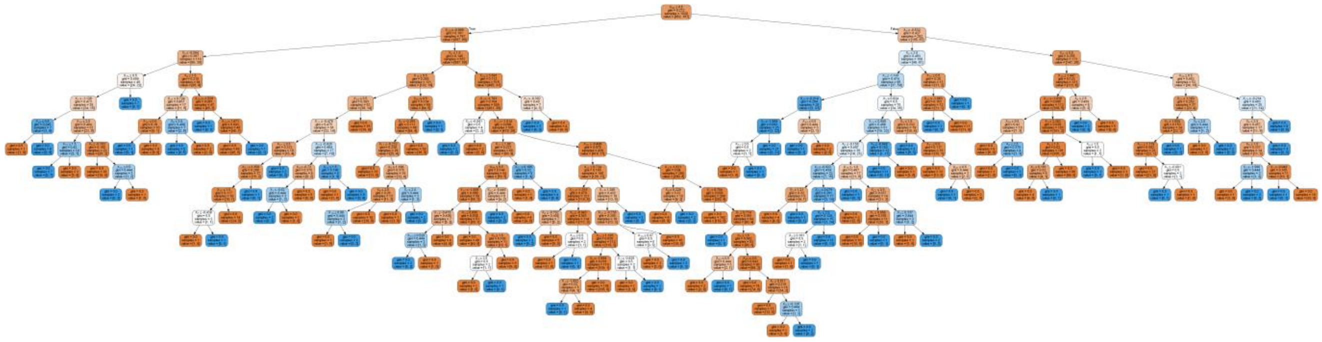
<u>Naive Bayes</u>

Impact on Naive Bayes was positive: computation time declined and accuracy increased. The performance of Naive Bayes classifier can be improved using preprocessing techniques, of which PCA is one. Correlated features can have a negative impact on Naive Bayes. PCA improves both accuracy and computational efficiency of Naive Bayes, as PCA tries to find uncorrelated variables and (depending upon the variation that needs to be retained) reduces number of features.

<u>Decision Tree</u>

Impact of PCA on decision tree was negative: both computation time and accuracy suffered. Though the number of nodes did not change by huge amount (Before PCA: 237 & After PCA: 227), the depth of the tree increased quite visibly (Before PCA:14 & After PCA:23).

```
Decision Tree without PCA
Nodes: 237
Depth: 14
```



```
Decision Tree after PCA
Nodes: 227
Depth: 23
```