PROJECT REPORT

ON

HUMAN RESOURCE ANALYTICS

kaggle™

SUBMITTED BY

Ajay Saini

Brief Summary: My approach for this challenge started with understanding the problem from business point of view and identify the possible reasons for the problem they are facing. After this I wanted to become more familiar with the provided dataset and find out which variables are most important for our analysis. I looked for relationships among variables, find missing values and outliers and generate new features. This was the most important step and I came back to this step several times during modeling the data. For modeling, I implemented 7 classifiers and compared their results. I also performed ensembling by running all the models for 3 times and then averaging their score, which marginally increased the accuracy.

Below I have provided a detailed description of my approach and what other steps I would have taken for improvement.

Detailed Description:

Step 1: Understanding problem and generate a hypothesis

Client: Anonymized company

Problem: Their best and most experienced employees are leaving the company prematurely. So, they want to predict which employees would be the next to leave and then try to retain them.

The possible reasons an employee might leave a company are low salary, extra work hours, very less or more than enough work, bad work culture, no promotion, unchallenging work, inflexible schedule, etc.

Data: A data set containing various features related to employees is provided.

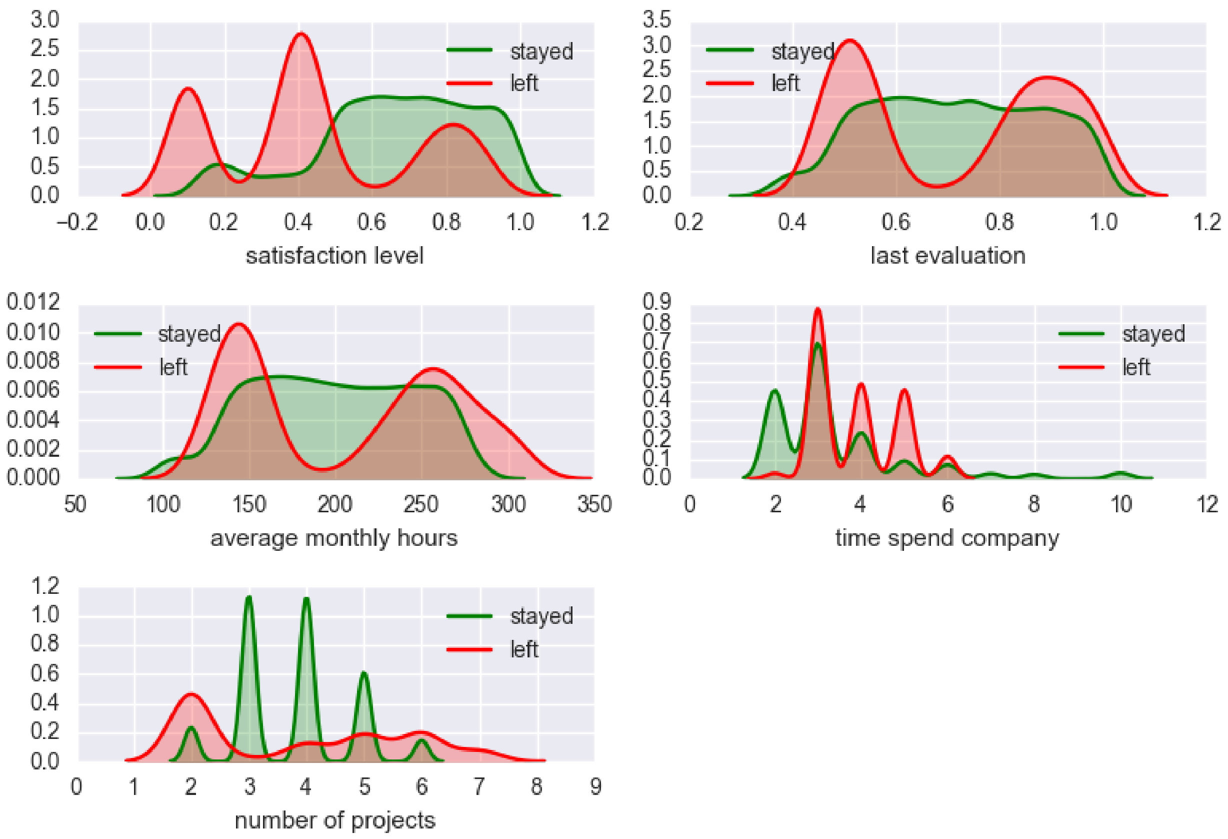
Step 2: Data Exploration

Target variable: 'left' having 2 outcomes i.e. 0 and 1. So, it's a classification problem and to predict it we need to find parameters that will be highly correlated with it.

No missing data – Great!

I begin by comparing the density of the continuous variables among those who left and those that have not. There are five variables to consider:

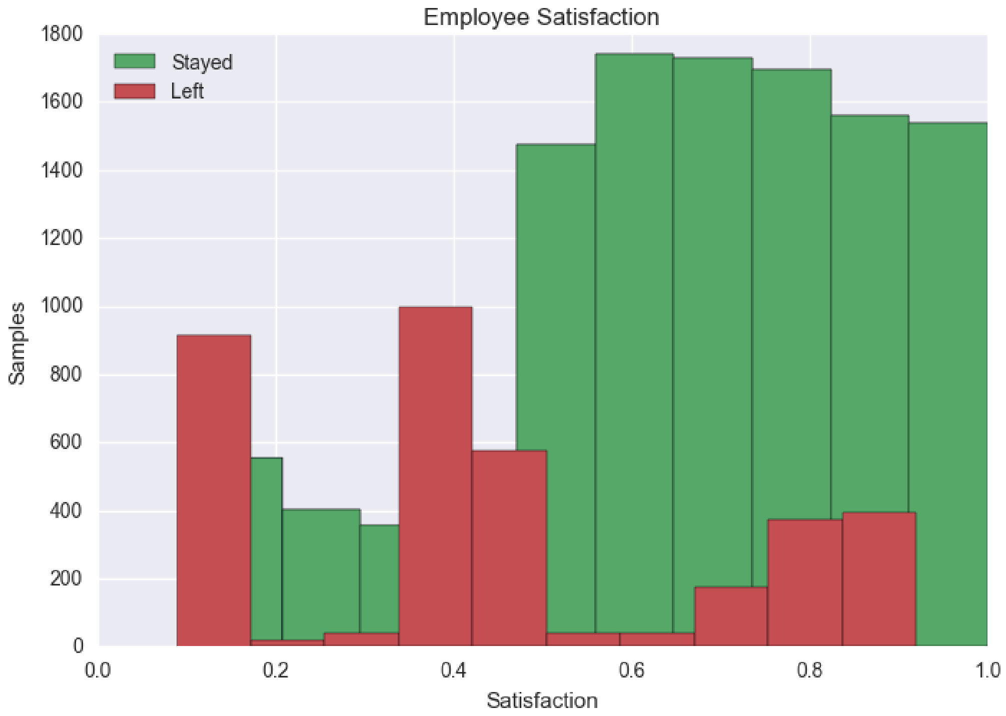
1. Satisfaction level
2. Last evaluation
3. Average monthly hours
4. Time spent at company
5. Number of projects

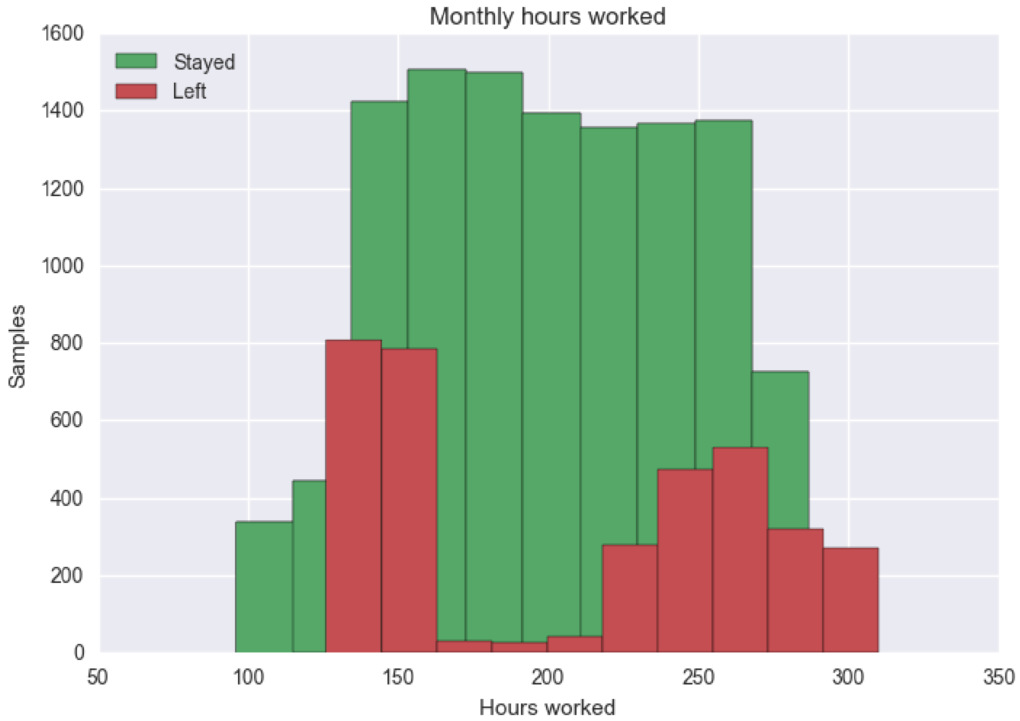


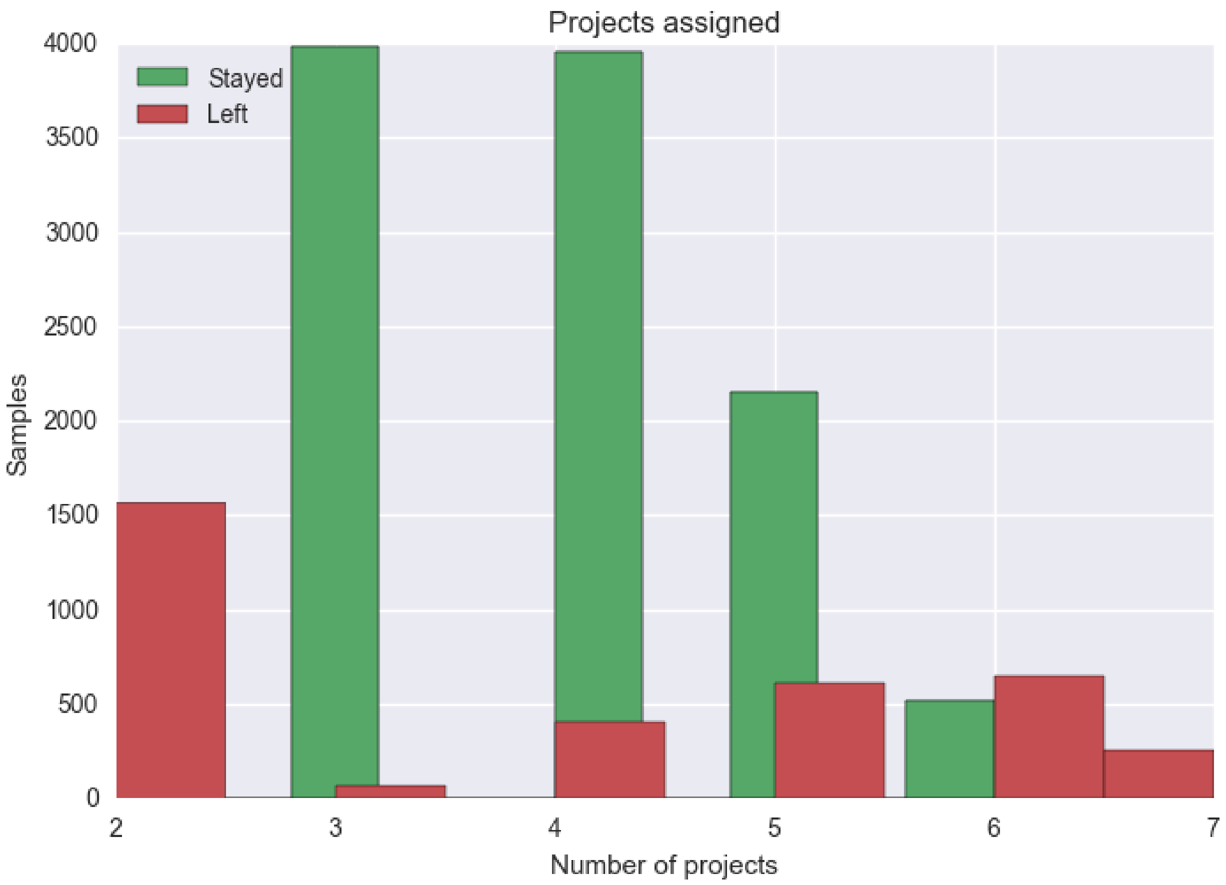
From the plots above, we can see that the relationships are not linear.

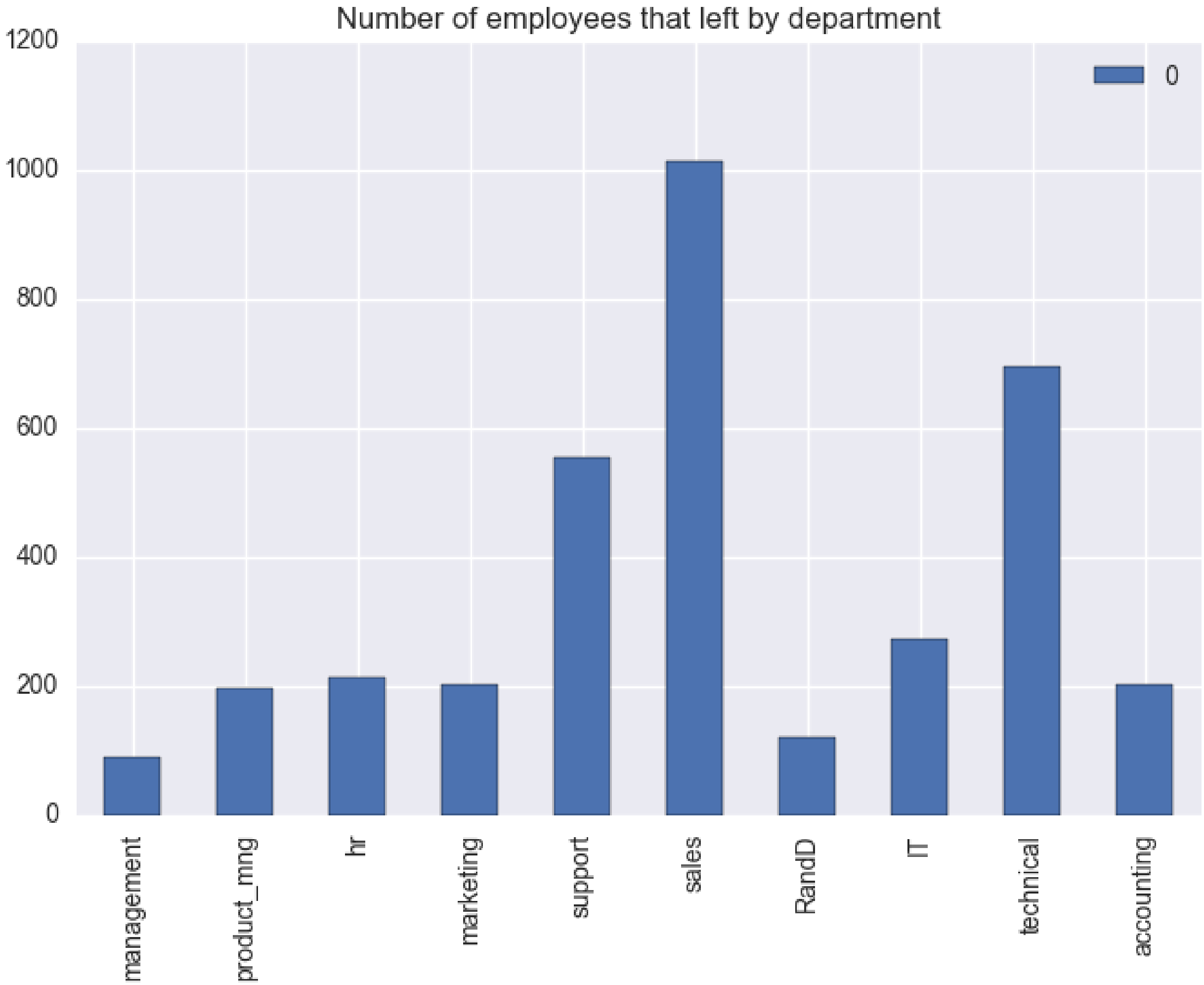
- Those with low satisfaction levels are more likely to leave. However, there is an increase in the density at a satisfaction level of around 0.8.
- Those who worked either a small number of hours (~140) or a very large number of hours (~260) are more likely to leave.
- Those with low evaluation and those with high evaluations have more chances of leaving.
- Those spending 3 years or more in the company are more likely to leave.

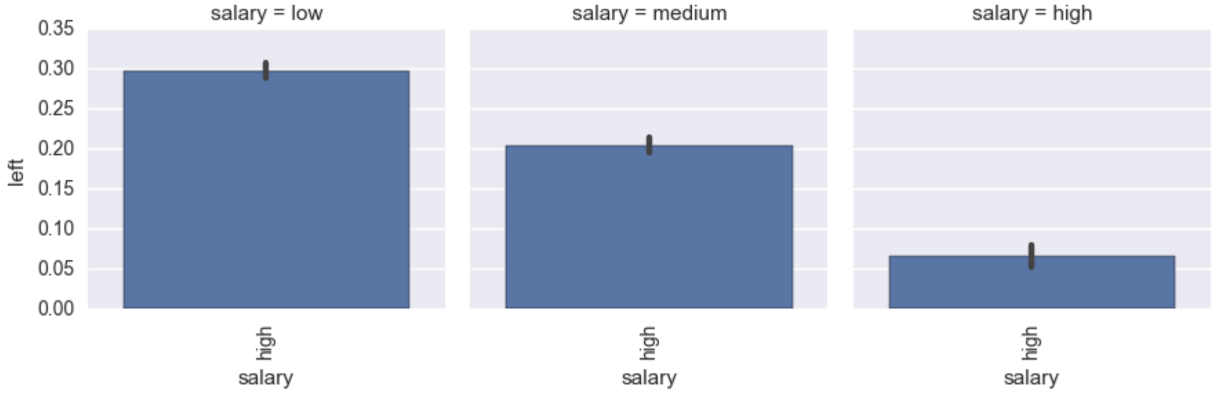
Next, I explored the number of employees left by each variable as following:

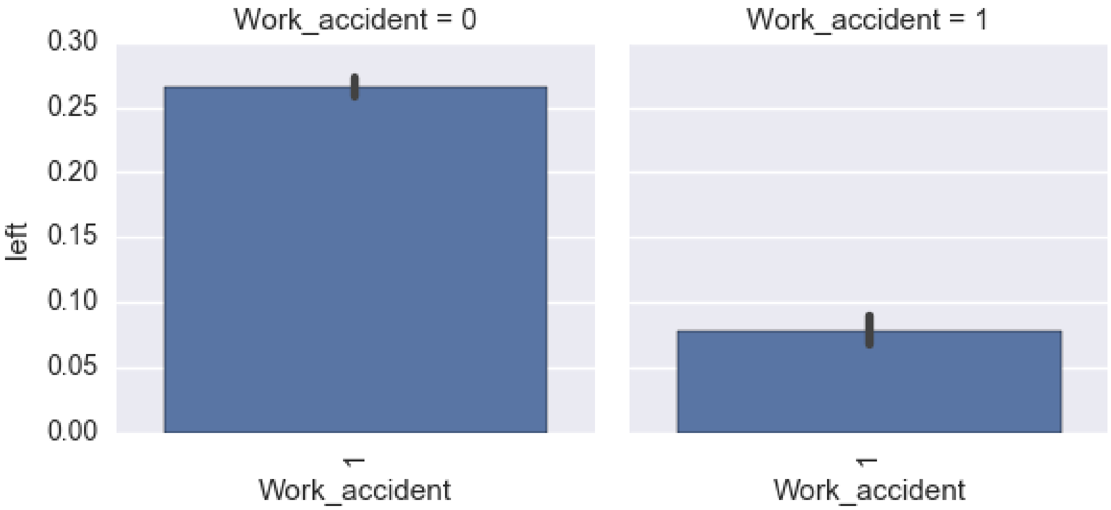

Employee Satisfaction
🟩 Stayed
🟥 Left
Samples
1800
1600
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Satisfaction

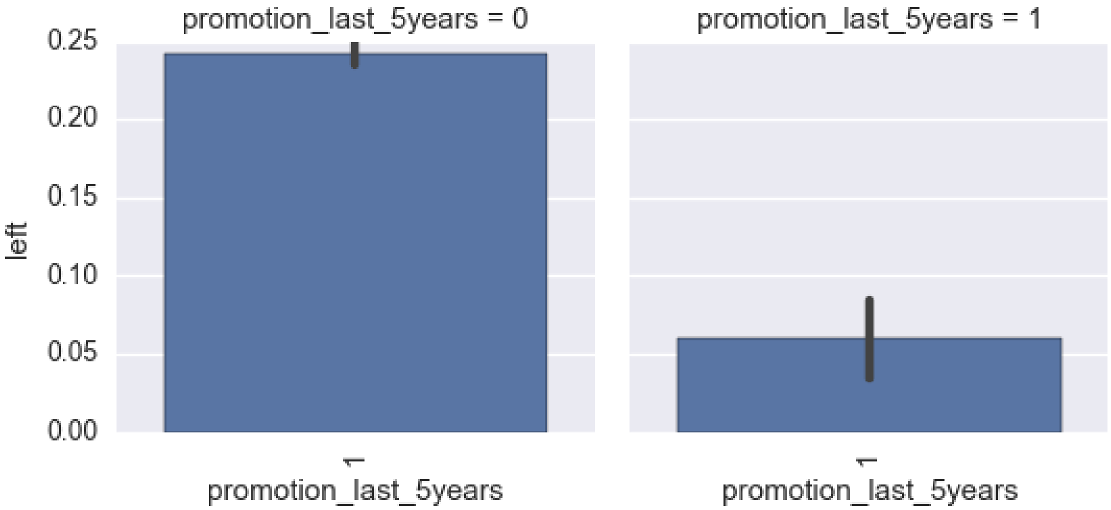

Monthly hours worked
🟩 Stayed
🟥 Left
Samples
1600
1400
1200
1000
800
600
400
200
0
50
100
150
200
250
300
350
Hours worked

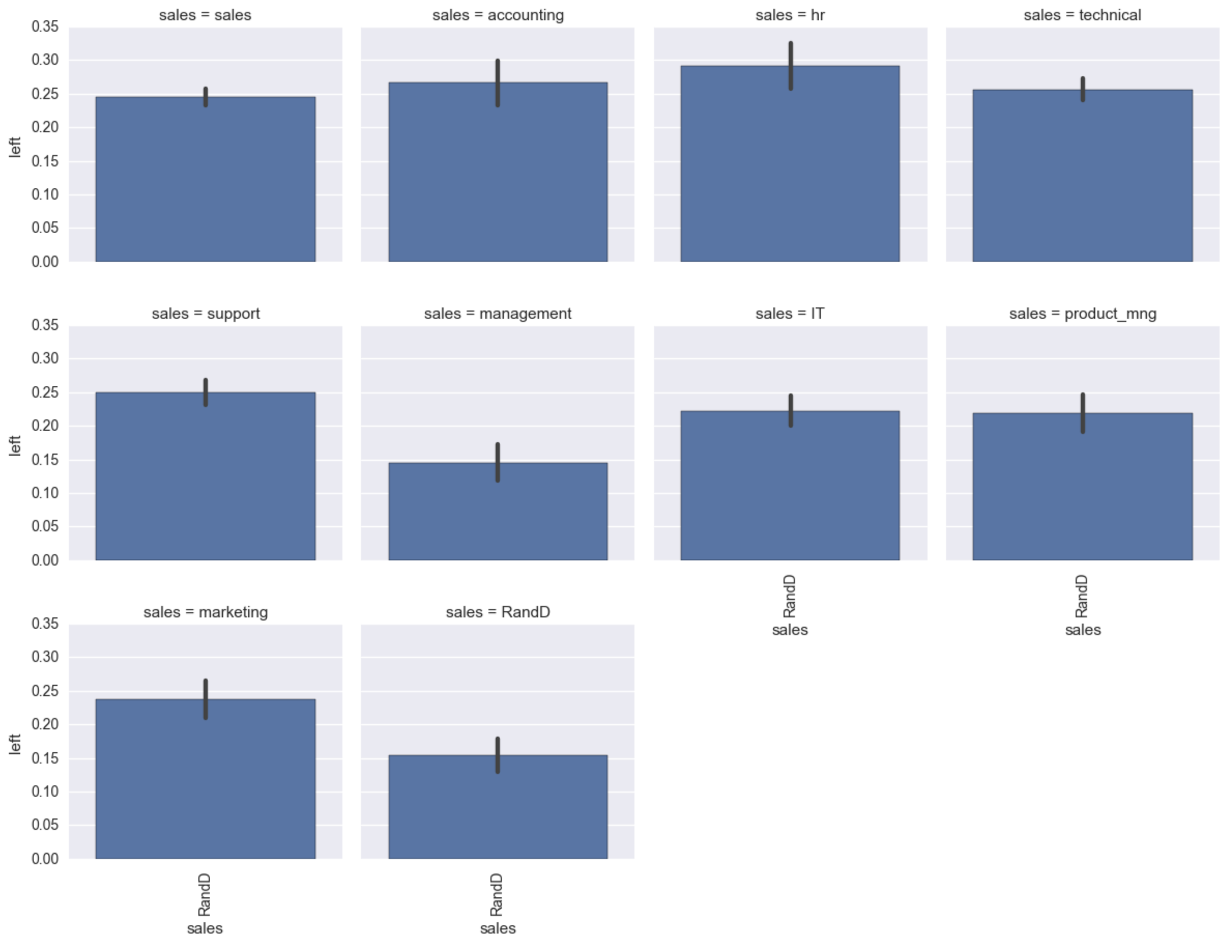

Projects assigned
🟩 Stayed
🟥 Left
Samples
0
500
1000
1500
2000
2500
3000
3500
4000
2
3
4
5
6
7
Number of projects


Number of employees that left by department
1200
1000
800
600
400
200
0
0
management
product_mng
hr
marketing
support
sales
RandD
IT
technical
accounting


salary = low
salary = medium
salary = high
left
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
high
salary
high
salary
high
salary


Work_accident = 0
Work_accident = 1
left
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
Work_accident
1
Work_accident


promotion_last_5years = 0
promotion_last_5years = 1
left
0.25
0.20
0.15
0.10
0.05
0.00
1
promotion_last_5years
1
promotion_last_5years



From the above plots, we can conclude that:

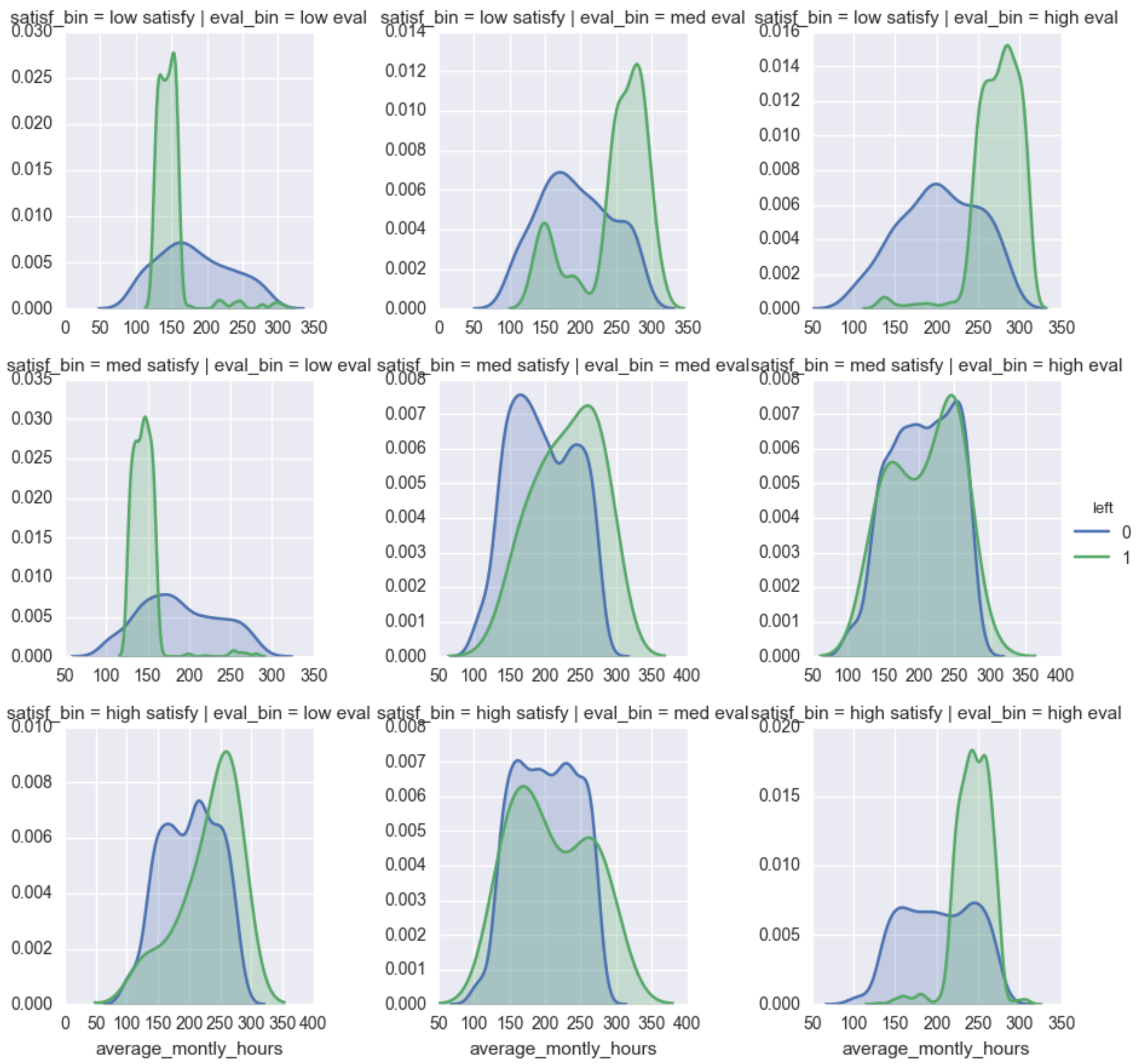
1. Low salary workers are more likely to leave
2. Those who had a work accident are less likely to leave
3. Those who were promoted in the past 5 years are less likely to leave
4. Those in management and R&D are less likely to leave

Step 3: Data Preparation and Feature Engineering:

Now, I wanted to convert the ‘satisfaction_level’ and ‘last_evaluation’ into categorical variables because the categorical variables allow you to capture much more complicated relationships rather than continuous variables which estimates the linear component of a relationship.

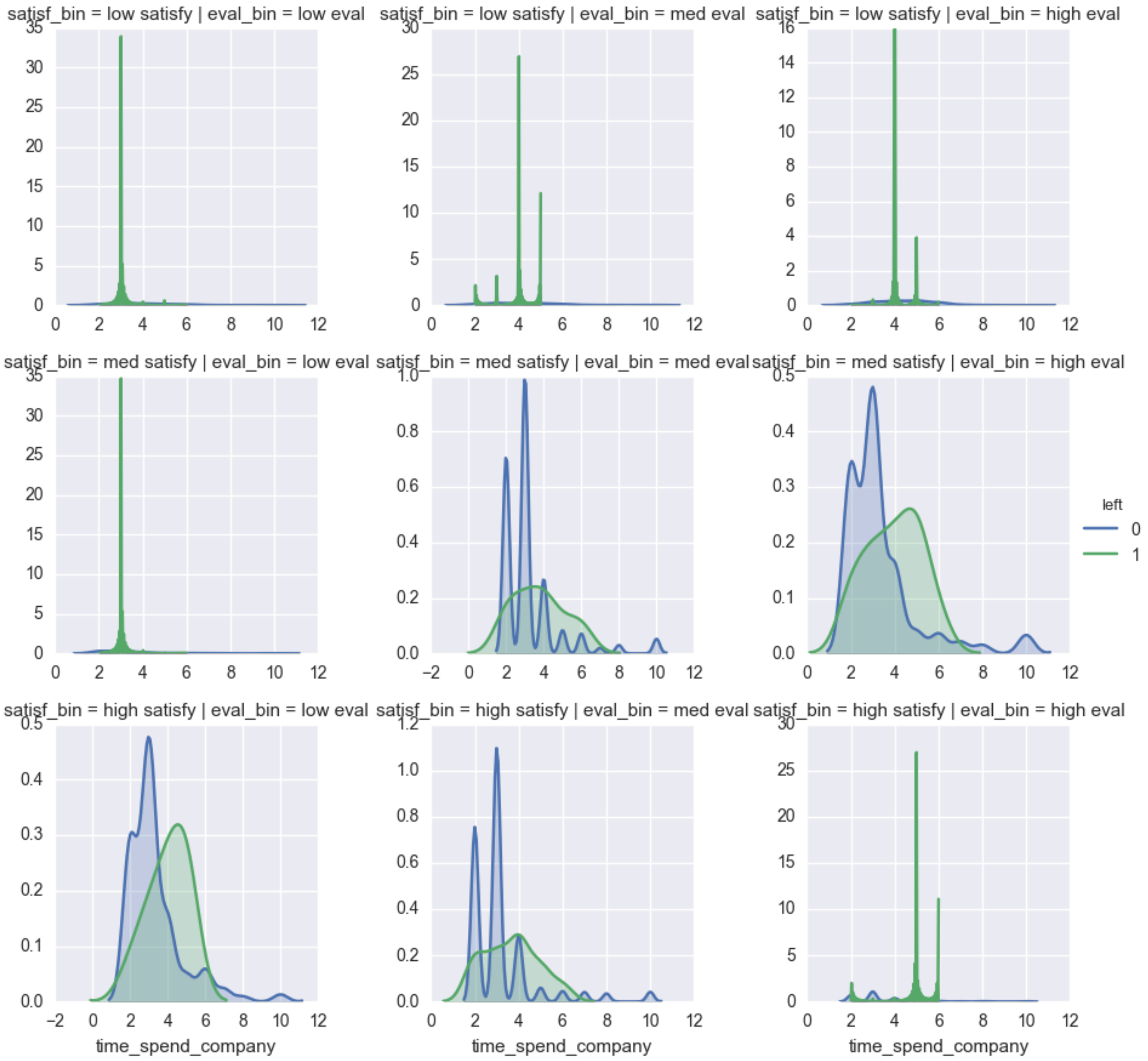
So, I performed binning on these variables by creating 3 bins for each having low, med and high values.

Next, I wanted to see the relationship between each of new variables created by performing binning.

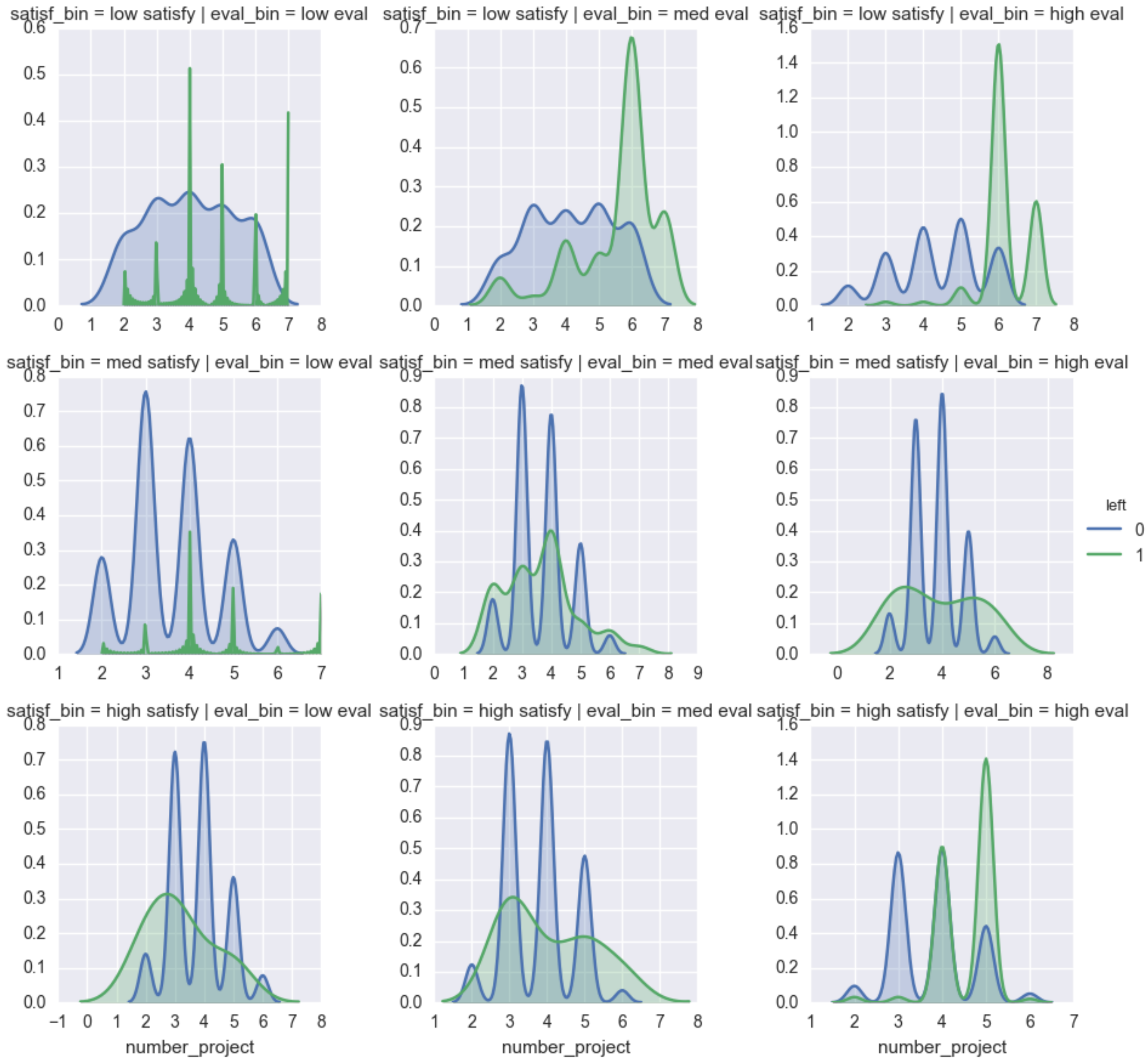


From the above plot we can conclude that:

1. low/med satisfaction and low evaluation: those working less hours are more likely to leave. The pattern changes for high satisfaction, where the distributions overlap and those who worked more hours were more likely to leave
2. Those with low satisfaction and med/high evaluations were more likely to leave if they worked for many hours
3. Those with high evaluation and high satisfaction were more likely to leave if they worked more hours.



This graph doesn't provide us much insights as there is an overlap for most of the categories.



For number of projects, although the distributions overlap. But those with low satisfaction and med/high evaluation and high satisfaction with high evaluation were more likely to leave if they had a large number of projects.

These results suggest that separating the evaluation and satisfaction into categories and interacting them with hours spend and number of projects could improve the predictions.

So, I created 5 new variables as following:

Var 1 = high satisfy * high eval * average_montly_hours

Var 2 = low satisfy * med eval * average_montly_hours

Var 3 = low satisfy * high eval * average_montly_hours

Var 4 = low satisfy * low eval * average_montly_hours

Var 5 = med satisfy * low eval * average_montly_hours

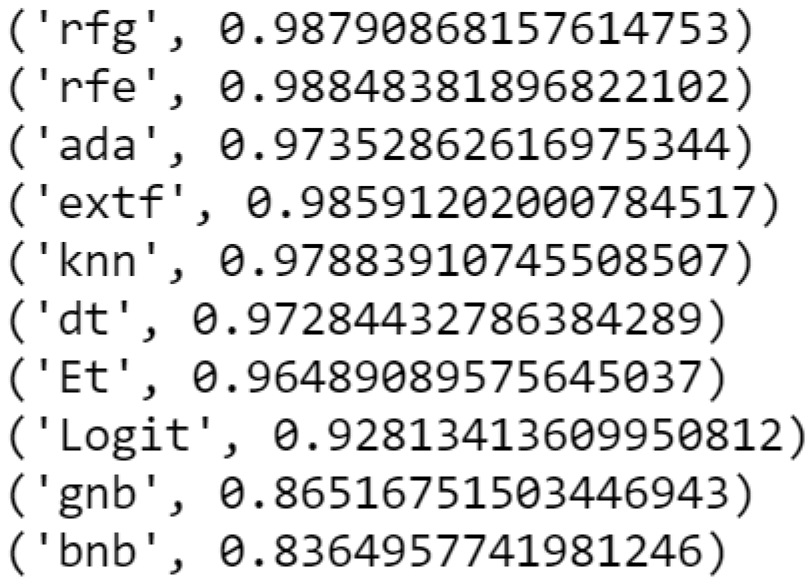
Now, let’s build model on this dataset and then we will see if more feature engineering is needed.

Step 4: Model Building

I implemented following 7 classifiers most popular for binary classification:

1. Random Forest
2. AdaBoost
3. k Nearest Neighbor
4. Decision Tree
5. Logistic Regression
6. Gaussian naïve bayes
7. Bernoulli naïve bayes

I ran the all the classifiers for 1 time and got the following results:

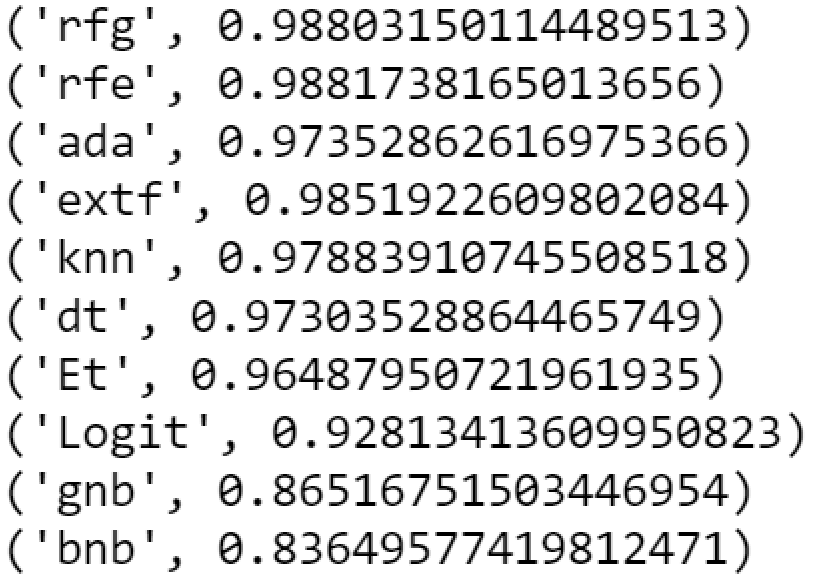


```
('rfg', 0.98790868157614753)
('rfe', 0.98848381896822102)
('ada', 0.97352862616975344)
('extf', 0.98591202000784517)
('knn', 0.97883910745508507)
('dt', 0.97284432786384289)
('Et', 0.96489089575645037)
('Logit', 0.92813413609950812)
('gnb', 0.86516751503446943)
('bnb', 0.836495774198124 6)
```

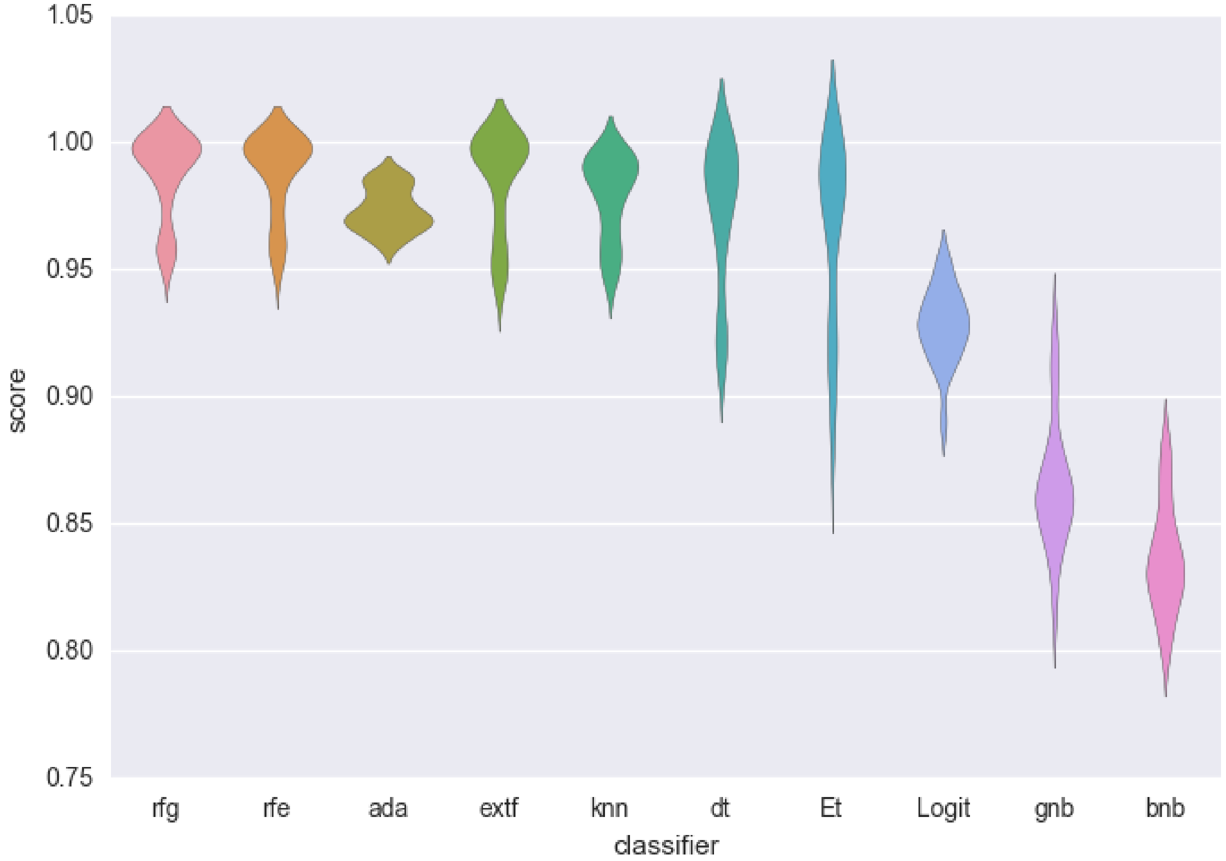
Step 5: Ensembling

For ensembling, I just used the basic version of combining the results of same classifier. I ran all the classifiers for 3 times and then took an average of their scores. As expected it increased the results

For n= 3:



```
('rfg', 0.98803150114489513)
('rfe', 0.9881738165013656)
('ada', 0.97352862616975366)
('extf', 0.9851922609802084)
('knn', 0.97883910745508518)
('dt', 0.97303528864465749)
('Et', 0.96487950721961935)
('Logit', 0.92813413609950823)
('gnb', 0.86516751503446954)
('bnb', 0.83649577419812471)
```



Although these results are very good on this data but may change for out-of-box data. So, a better approach for ensembling could have been to build different models using different parameters and then average their results.

RESULT:

Most of the classifiers have good performance. I'll choose Random Forest as our classifier of choice because it has high ROCAUC and low variation in the CV. Also, it is it is fast and fairly insensitive to tuning.

Additional steps for improvement:

1. We can perform more statistical operations on the data and search for any other hidden pattern inside the data. Enriching the dataset by combining data from other sources is also an option.
2. Tune parameters for the classifiers.
3. We can also use Stacking on the above-mentioned classifiers to see if there is any further improvement. Remember to use out-of-fold predictions, else there will be serious over-fitting.

General Approach for Data Exploration

Data exploration is the most important step in a data analysis process. A significant amount of time should be spent on exploration and analyzing data, once you have got your business hypothesis ready.

Step 1: First, identify Predictor (Input) and Target (output) variables. Next, identify the data type and category of the variables.

Step 2: Explore the variables one by one. For continuous variables, we need to understand central tendency and spread of the variable using Boxplots and Histograms. For categorical variables, we can use Bar Charts to understand distribution of each category.

Step 3: Next, we need to find out the relationship between two or more variables. This includes finding correlation among variables using Scatter Plots and performing statistical tests such as t-test, ANOVA, chi-square test.

Step 4: Now comes the most important step i.e. Missing Value Treatment. Missing data can lead to a biased model and reduce the fit of the model. There are several ways to treat the missing values depending on certain conditions:

i. List-wise Deletion
ii. Pair-wise Deletion
iii. Mean/Median/Mode Imputation
iv. kNN Imputation
v. Creating Prediction Model to estimate missing values from available values

Step 5: Outlier Treatment is as equally important as treating missing values. Outliers can be detected using boxplots, scatter plots and histograms. Generally, data points that are 3 or more standard deviations away from mean (i.e. outside 5th and 95th percentile) are considered outliers. There are several ways to treat the missing values depending on certain conditions:

i. Deleting Observations
ii. Mean/ Median/Mode Imputation
iii. Transforming and binning the values
iv. Treating Separately

Step 6: The art of extracting more information from existing data is what makes a Data Scientist differ from a Data Analyst. Although, feature engineering is the final step in data exploration but you will need to come back again and again to this step to refine your model. This step involves two parts i.e. Variable Transformation and Feature Creation. Variable Transformation is needed when we want to change the scale of variables or transform a complex non-linear relationship into linear relationship. It can be performed by taking the log, square/cube root or binning the values. Creating new features includes methods like converting categorical variables into dummy numerical variables.

General Approach for Data Analysis

Step 1: The very first thing I would do is clearly define the problem or objective so I have a solid direction. We need to understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition.

Step 2: The next step is to understand and explore the dataset. This involves steps such as Variable Identification, Uni-variate and Bi-variate Analysis, Missing value treatment, Outlier detection, Variable Transformation and Feature Creation. We can also enrich our dataset with other data sources.

Step 3: After processing the data, we are now ready to build model. In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. We often need to step back to the data preparation phase based on the performance and results. Ensembling and Stacking can be applied on the top classifiers. We should keep modeling the data until we find most significant and valuable results.

Step 4: Before proceeding to final deployment, the model should be thoroughly evaluated and it should be made sure that it properly achieves the business objectives.

Step 5: Lastly, I would implement the model and track my results.

In my opinion the first two steps are the most important steps in a data analysis process. If you are not clear with the project requirements, then you are just wasting your time. And certainly, good data is more worth than a good algorithm. So, more time should be spent on feature engineering rather than trying bunch of different algorithms to get good results.