**Machine Learning Engineer Nanodegree**

**Capstone Project**

Arshad Khurshid

11/11/2017

**Definition**

**Project Overview**

Machine Learning Engineer Nanodegree course has emphasized on different type of Supervised and Unsupervised Learning. Supervised Learning deals with Regression and Classification. Classification is being implemented across all the fields for different applications and are very prominent in Cyber Security and Medical diagnosis. With boom in data, Classification model has gained momentum to understand the cause of change in behavior in consumer. This helps business to work on data and not to take decision just on intuition.

A company invest lot on employee to train them and make them ready for next generation business. Once you invest in skill enhancement of an employee you need to use it for benefit of business. Employee may be agitated even if they are being paid well as human have aspiration and if aspiration is fulfilled then they perform to their maximum capability. I was looking for dataset which has features that can be reason for employee to quit and Kaggle gave me one. Retaining an employee means retaining knowledge and they are the one who grooms the people working in one level down thus helping to increase and knowledge base for all.

Dataset Link: <https://www.kaggle.com/ludobenistant/hr-analytics>

**Problem Statement**

Managing people in a company is challenging job. When it comes to attrition HR has a tough job to identify the root cause of attrition. Most of the time it happens that diagnosis based on intuition is wrong and hence the policy implemented has more damaging effect and situation worsen. What if HR knows in that particular employee is more probable of leaving and could take damage control action. What is they know why employee are leaving and features that is having more damaging effect on them. Since I need to emphasize on 2 groups Classification of employee into Left and Stayed and I treat it as Classification problem. The other is the reason why people are leaving. For this i will drop the decision feature (left) and will group features in different nodes and find the features that has most impact.

**Features and description**

* satisfaction\_level : Level of Satisfaction
* last\_evaluation : Time since Last performance Evaluation
* number\_project : Number of Project completed while at work
* average\_montly\_hours : Average monthly hours at workplace
* time\_spend\_company : Number of years spent in the company
* Work\_accident : Whether the employee had a workplace accident
* left : Whether employee left the workplace or not
* promotion\_last\_5years: Whether employee was promoted in last 5 years
* sales : Department they work for
* Salary : Relative level of Salary(high)

Model will be predicting if person is potential employee who will stay or leave the company. Satisfaction level is important attribute and a person not satisfied will probably leave the organization. Last evaluation means that performing person is not getting enough feedback and hence enthusiasm to continue work is lost. Number of project completed does not seem important attribute as project can be small or big. Some employee may have worked on big projects continued for long time whereas some may work for high number of small projects. Average monthly hours plays important role as employee putting too much time for work may have no time to relax and hence will burn out and would like to move on. Work accident plays important role as it concern with security of people and that may require people to quit. Left is the decision column which will used to predict. Promotion in last 5 years and salary should have same impact and hence I believe salary should be used and promotion to be dropped. Sales is department value and will be interesting to see which department is more impacted. But I do not think it impacts on decision. Salary an important column for employee to quit.

The dataset has total of 14999 rows and 10 columns. I will read the csv file using pandas library. Will analyze the data to check for any null values in any of columns and impact of each column on output. With this model i hope to analyze the employees who are more likely to exit and could do damage control exercise to retain them thus retaining talent and boosting company performance. Target variable (left) is imbalance dataset. It contains 3751 records of employee who have left the company. 11428 records of employee who stayed in the company. Here we are trying to predict employee who can quit. Dataset has been extracted from Kaggle.

Dataset Link: <https://www.kaggle.com/ludobenistant/hr-analytics>

Company is in a situation where its talented and experience employee are quitting jobs and there might be n number of reasons. HR job is to identify the root cause and take a remedy action. Machine Learning model will help them to identify the people who could quit in future and the reason for quitting the jobs. So that a preventive measure is taken and employee could be retained. I plan to use Classification model of Supervised Learning techniques to learn from history data and when existing employee is pass to the model it could predict if employee will be moving on and damage control action will be taken. Couple of step will be taken for preprocessing of data. Will check for null value for feature. If available will fill with mean for numerical column and mode value for categorical column. Will do one hot encoding for categorical column. There are total of 10 features available. Since all of them will not have equal impact on prediction I will use PCA to identify the most predictive feature and use those. I will use Ensemble machine learning model (AdaBoost /Gradient Boosting) to understand the employee moving on. Will finalize any one based on performance. I will also use Stochastic Gradient Descent Classifier (SGDC) and SVM to check their f score and if performing better than Ensemble then will use it in fine tune and create model. Since the goal of model is to understand why employees are moving on I will consider more interpretable model than highest prediction accuracy.

**Metrics**

Classification model will be evaluated on F Beta score and Accuracy, precision and recall. We can use F-beta score as a metric that considers both precision and recall:

*Fβ* = (1+*β*2) ⋅ *precision* ⋅ *recall / (*(*β*2⋅*precision*) + *recall)*

In particular, when *β*=0.5, more emphasis is placed on precision. This is called the F0.5 score (or F-score for simplicity).

Accuracy measures how often the classifier makes the correct prediction. It is the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

Precision tells us what proportion of messages we classified as left, actually left. It is a ratio of true positives(predicted value for person left and actually person left) to all positives (all the person left company irrespective of true or false), in other words it is the ratio of   
[True Positives/(True Positives + False Positives)]

Recall (sensitivity) tells us what proportion of employee that left were classified by us as left. It is a ratio of true positives (employee that left, and employee actually left) to all the employee left, in other words it is the ratio of [True Positives/ (True Positives + False Negatives)]

**Analysis**

**Data Exploration**

I have calculated total number of records available in dataset. Number of people moved out and number of people stayed. Percent of employees moved on. Total number of features available. To find null values in dataset and datatype of imported data.

By looking at the data type we can understand that there are two columns sales and Salary which is categorical column and rest other are numerical columns. Categorical column need to be one hot encoder.

Total number of records: 14999

Total number of employee who have moved on: 3571

Total number of employee who have stayed: 11428

Percent of employee who have moved on: 23.81%

Total number of null rows in Dataframe: 0

Datatype as below

satisfaction\_level float64

last\_evaluation float64

number\_project int64

average\_montly\_hours int64

time\_spend\_company int64

Work\_accident int64

left int64

promotion\_last\_5years int64

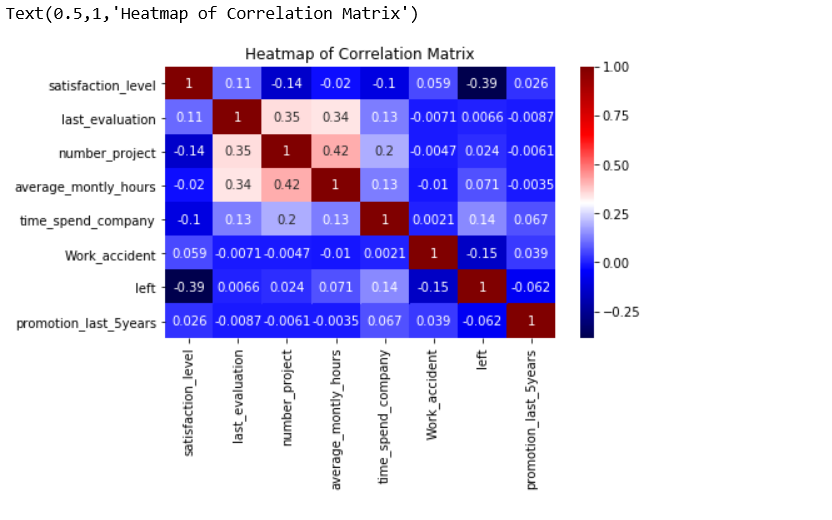
sales object

salary object

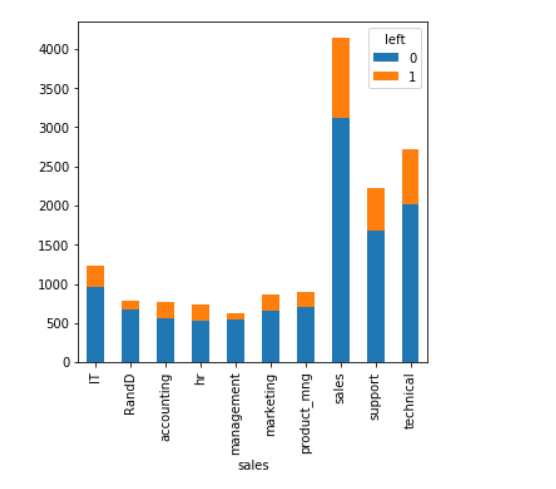
dtype: object

**Exploratory Visualization**

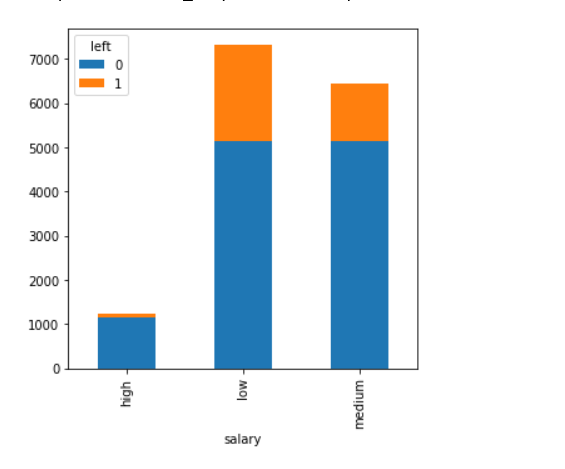
I have first generated Heatmap of correlation matrix understand the relation of each column on the other column. It states that average monthly hours and number of projects are highly correlated and hence using any one column for machine learning model should good.



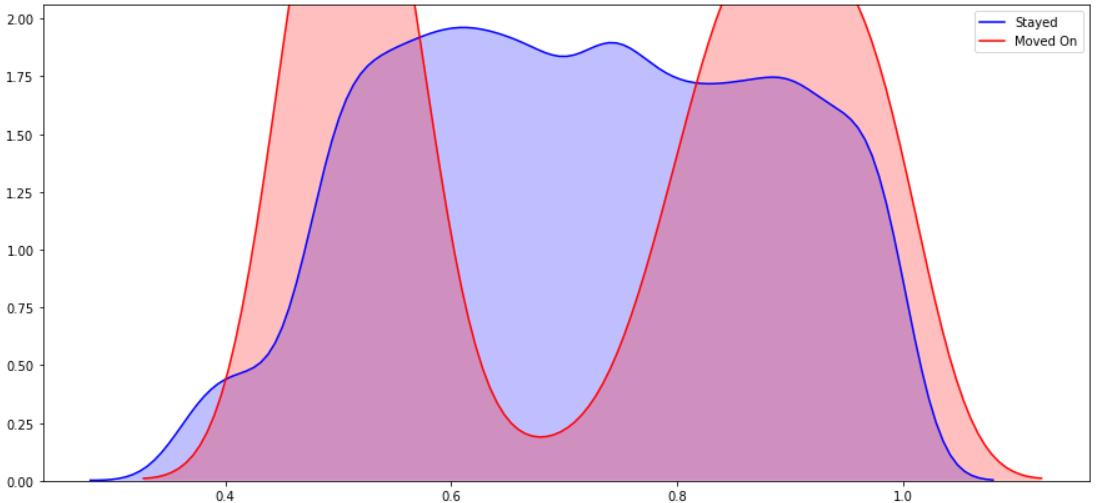
I would like to find out which department are most affected to understand if it is departmental problem or attrition is across the board. Below graphs confirms which department is most impacted. One of features to look into it.



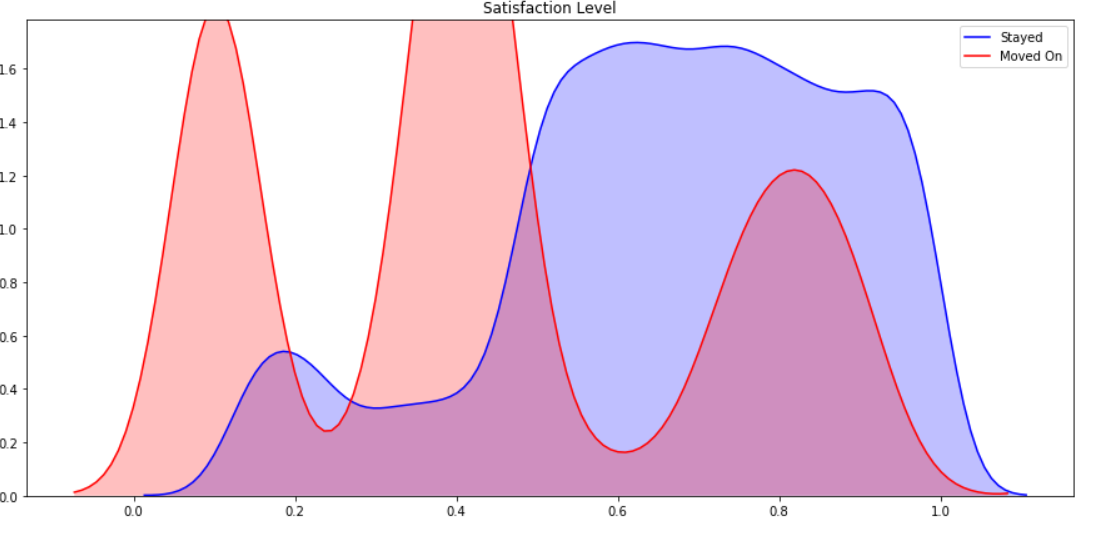
As per my intuition I want to check if people are low paid is impacted and will like to move on for better opportunity and below graphs confirms that most people with low salary are one to quit. Moderate number of median salary have quit job and very few with high salary.



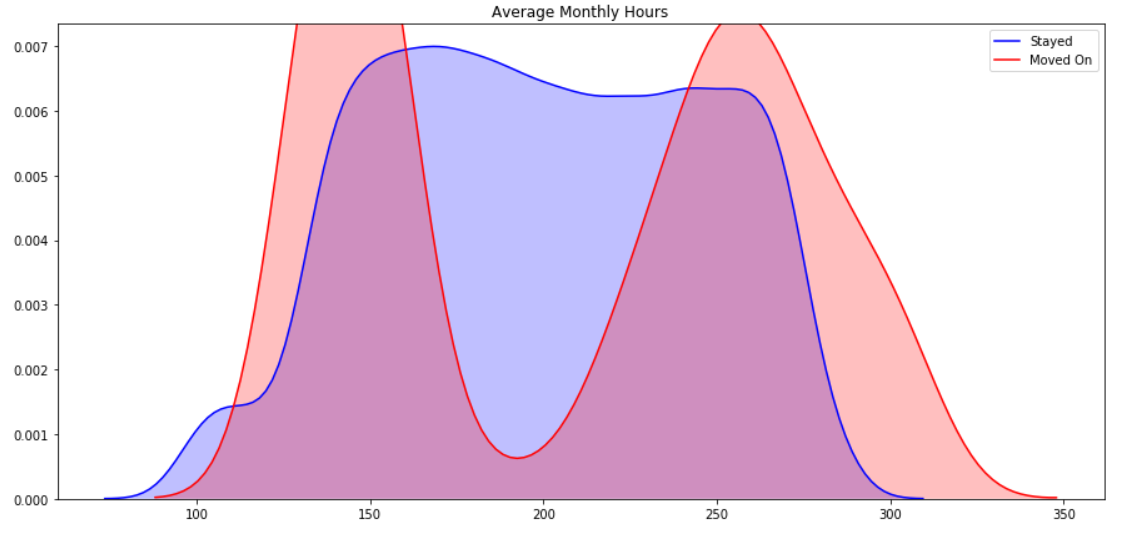
I tried to plot graph on last evaluation and left column to understand the impact of last evaluation on decision to stay in a company and it was found that evaluation score between 0.4 to 0.6 and 0.8 to 1.0 were the one leave company. Giving us view that one with mid-level of evaluation usually stays.



I tried to plot graph with satisfaction level and left column to check what is people are not satisfied. Do they make decision to move on and graphs prove that indeed they are one to move out.



Another column which could impact on people was average monthly hours in company and guess was that if a person spends too much time he may be burn out and would quit. This is what shown in below graph.



**Algorithm and Techniques**

One of the challenge to train a model is to bring all the data elements to same scale as different scale will unnecessary give more weight to the feature not required. For example if we take 2 features of a person such as height and weight to identify if person is male or female. Height measure in feet will vary between 4-7 feet but weight will vary from 100 - 300 pounds. Thus small change in weight will have high impact on decision making.

In this project numerical column identified has been brought to one scale where values lies between 0 and 1.

Categorical column has been converted to numerical column with value 0 and 1. All values has been converted to columns.

Data now converted to training set and test set in 80:20 ratio.

For training we have variable as X\_train and y\_train.

For testing set we have data X\_test and y\_test.

**Benchmark**

As a benchmark model I have used Logistic Regression classification model and Confusion matrix, accuracy and F Beta score is calculated so that any improvement will add to the performance of model.

F Beta will be used for measuring performance, Accuracy is for support purpose only.

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | ACTUAL | |
| PREDICTED | Left | Stayed |
| Left | 227 | 474 |
| Stayed | 165 | 2134 |

Accuracy: 0.787

F Beta Score: 0.500