MACHINE LEARNING REPORT

SOHAMJIT MUKHJEE

WWW.SOHAMJIT.MUKHEREE@STUDENTS.IE.EDU

Machine Learning Assignment 1 – Sohamjit Mukherjee

Entire Process Overview:

Quick Overview		
Step	**Process**	**Conclusion**
Data Preparation	Summary Statistics	9 columns in the dataset and the dat set is balanced
	Missing Value	There are no missing values in the data set
	Box plot	Revelled there are no major outliers
	Refactor to Data Type	Categorical variable are converted to factor data type
EDA	Correlation Matrix	Discussed in detail in report
	Relation among employees leaving and department	People R&D and management are less likely to leave
	Satisfaction Level vs left	People less satisfied are more likely to leave. (Discussed in detail in report)
	Last Evaluation vs Left	No relationship was found in this case
	Salary vs Left	As expected, people with high salary generelly tends to stay in the company
	Distribution of Numerical variables	Distributions are checked to fix skewness latter
Baseline Model	Logistic Regression	Train : 75 % Test : 25 % Process of Separation: Random Sampling Accuracy : 78.5%
	XGBoost	Train : 75 % Test : 25 % Process of Seperation: Random Sampling Number of trees : 100 Accuracy : 90%
Feature Engineering	Feature Creation	average_monthly_hours_per_project = avg_monthly_hours / number_of_project total_time_spend_in_company = Avg_monthly_hours * time_spend_in_company IS_Department_R&DorManagement = if department is R&D or Management sat_eva = satisfaction_level * last_evalaution evaluation_per_project = last_evaluation / number_of_project
	Scaling	Scaling is done to make the unit of observation same
Intermediary Modelling	Logistic Regression	Train : 75 % Test : 25 % Process of Seperation: Random Sampling Accuracy : 82%
More Feature Engineering	One Hot Encoding	Converted the salary field by one hot encoding
	Skewness Detection	Detect skewness in numerical fields
	Cube Root Transformation	Skewed columns are transformed using cube root transformation
Final Modelling	Logistic Regression with Boosting	Test & Train - 90% Holdout - 10% Process - cross validation Folds - 10 Grid search - Yes Selection of Model - Best by Accuracy Accuracy on Holdout dataset - 97.5 %
	Variable Importance Score	Calculated and plotted (Discussed in detail in Reports)

Problem Overview: The dataset provided contains explanatory variables of about 15 thousand employees of a large company. The goal of the case study is to model the probability of attrition of each employee as well as to understand which variables are most important and need to be addressed right away.

Machine Learning Assignment 1 – Sohamjit Mukherjee

The result obtained will be helpful for the management in order to understand what changes they should make to their workplace so that they can make their employees stay in the company for longer time.

Data Preparation & Exploration: The data set contained two different classes with '0' signifying the employees who have not left and '1' meaning employees who have left the company. The dataset seems to be ***quite balanced*** with 76% belonging to class '0 '(didn't left) and rest to class '1'. The general statistics of each and every column revealed that ***there were no missing values and none of the numeric columns has outliers***. This was further confirmed by the Box plots, which also revealed non-existence of outliers.

Exploratory Data Analysis: The correlation matrix was created as the first step of the data analysis. This matrix would help us to answer few questions. Studying the overall the matrix, we can conclude that none of the features are highly correlated.

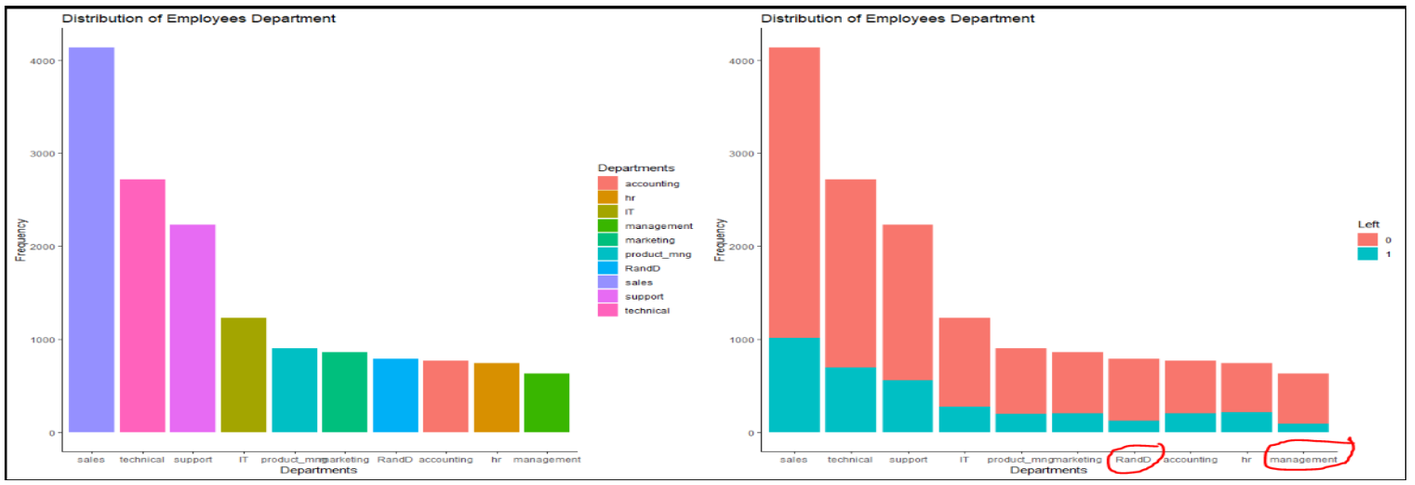
Still we can derive few insights from the correlation matrix:

- From the correlation matrix, we can see quite a high negative correlation between satisfaction level and those who left. Basically, people tends to leave if their satisfaction level is low.
- Secondly, people who work on many projects also have high average monthly hours and seems to get high evaluation ratings.
- Thirdly, even though it's a weak negative correlation but people having work related accidents tends to leave the company.

Further, it was found from the data set, that the people ***who works in R&D and in management are less like to leave*** compared to other. Also, as expected people ***having high salary do not leave the company usually***. But the most important trend was found in satisfaction level: ***People who have rated satisfaction level above 9.2 have never left the company and those who rated below 0 .1 have all left the company.***

Moreover, there seems to be three different clusters for people who left the company. It seems that a high proportion of people who voted 0.36 to 0.4 seems to leave the company.

Machine Learning Assignment 1 – Sohamjit Mukherjee



Baseline Model: For baseline model, the data set was simple split into 75% train and rest 25% as the test data set ***by simple random sampling.*** Two different algorithms ***(Logit & XGBoost)*** was run on the same train data set and the result was compared to the test data set. Having same train and test dataset also allowed us to compare the accuracy of these two algorithms. ***No advanced techniques, like feature engineering, cross validation or grid search had been used in this step.*** The **accuracy of the Logistic *regression was found to be around 78.5%*** on the test data set while ***XGBoost has an accuracy of about 90%*** on the test data set. Based, on this we can fairly conclude boosting performed way better than the normal logistic regression.

Accuracy: Logistic Regression (Baseline)

```
   Null deviance: 12371.0  on 11248  degrees of freedom
Residual deviance:  9633.8  on 11230  degrees of freedom
AIC: 9671.8

Number of Fisher Scoring iterations: 5

[1] "Accuracy 0.789333333333333"
```

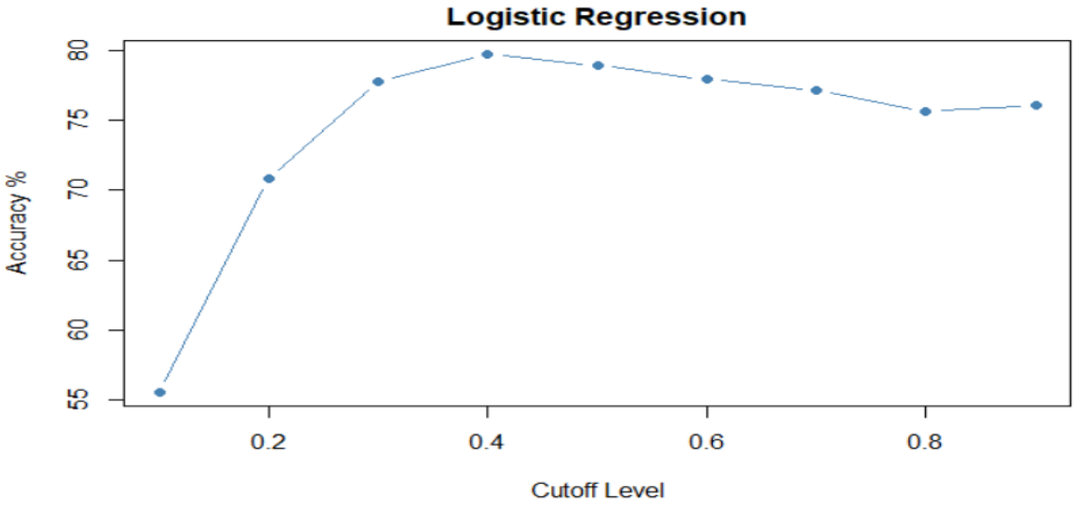
Error: XGBoost (Baseline). So, Accuracy = (1 – Error)

```
# Run the model
model = xgb.train(params = xgb_params,
                  data= train_matrix,
                  nrounds = 100,
                  print_every_n = 20L,
                  watchlist = watchlist)
```

```
[1]     train-merror:0.021157   test-merror:0.023200
[21]    train-merror:0.015468   test-merror:0.018933
[41]    train-merror:0.011201   test-merror:0.017600
[61]    train-merror:0.006756   test-merror:0.015467
[81]    train-merror:0.003911   test-merror:0.013333
[100]   train-merror:0.001778   test-merror:0.009867
```

Machine Learning Assignment 1 – Sohamjit Mukherjee

Also, the below shows the changes of accuracy over different cutt of level. ***The maximum accuracy occurs when cutoff is 0.4***



Even though there is not much difference in accuracy between cut-off value of 0.4 or 0.5.

Feature Engineering: The feature engineering steps was the most crucial steps of the process which helped us to increase our accuracy from 78% to close to 97.5%. ***5 new features are created from the existing columns in the data set.*** Most of the features created seems to play an important role in the final model, with the feature "sat_eva" & "total_time_spend_in_company" has a score of more than 60 out of 100.

Out than that, another steps which helped us to increase the accuracy of the model was detection of skewness and applying **cube root transformation** to fix the model.

Final Model: For the final model the data set was split into 90 -10. 90% of the data was used as train and test data set while the rest ***10% was kept out as holdout/ validation dataset.*** After repeated trials cross validation of 10 folds gave a higher accuracy while testing on d validation set. The hyper-parameter optimisation technique of grid search has been used in the final model to predict the best combination of hyper parameter.

```
train_control<- trainControl(method="cv", number  = 10, savePredictions =
"all"
                             ,selectionFunction ="best" ,search = "grid" )
```

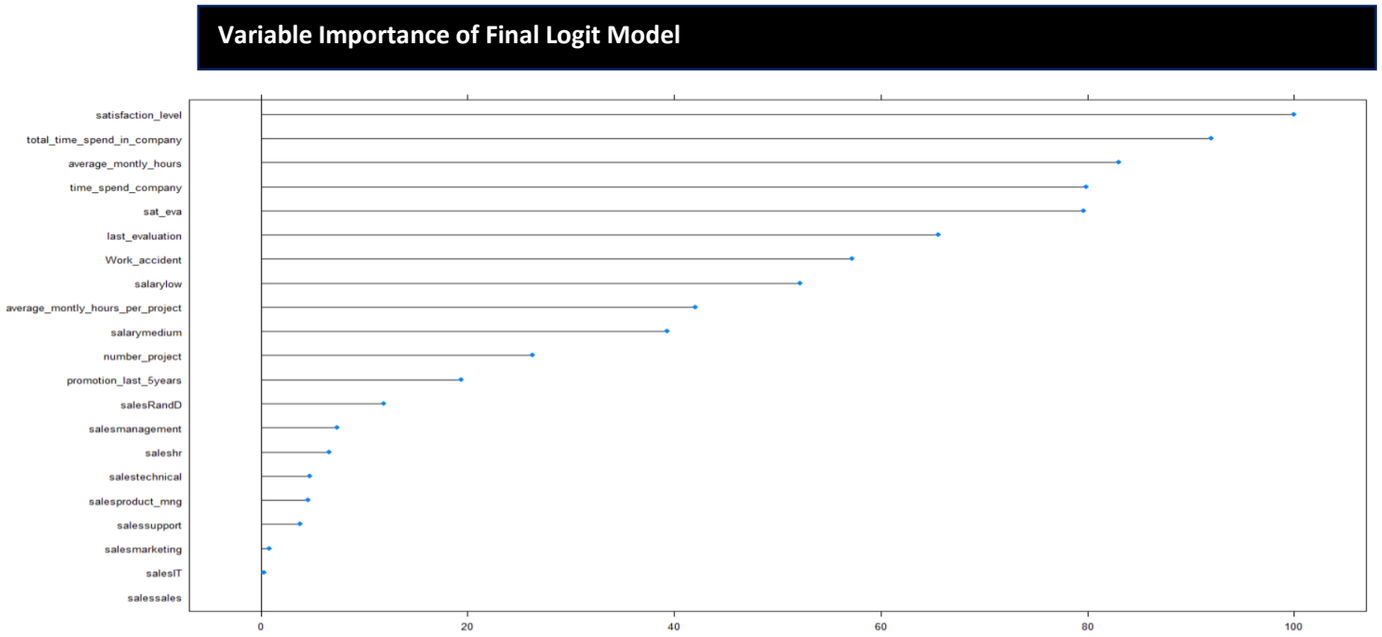
```
               Accuracy : 0.9746
                 95% CI : (0.9654, 0.982)
    No Information Rate : 0.7645
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9295
 Mcnemar's Test P-Value : 0.8711

            Sensitivity : 0.9843
            Specificity : 0.9433
         Pos Pred Value : 0.9826
         Neg Pred Value : 0.9487
             Prevalence : 0.7645
         Detection Rate : 0.7525
   Detection Prevalence : 0.7658
      Balanced Accuracy : 0.9638

       'Positive' Class : 0
```

Machine Learning Assignment 1 – Sohamjit Mukherjee

Below is how the final classification table of the holdout data set looks like. ***The sensitivity and specificity both quite high around 98.5 and 94.9% respectively.***

Classification Table (Holdout Data set)			
predictions	Reference		
	0	1	
	0	1128	20
	1	18	333

Variable of Importance: From the variable importance point of view we can say that ***satisfaction level, total time spend in company and average monthly hours*** are the three mostly important feature. ***Most of the new variables created through feature engineering like sat_eva, & total_time_spend_in_company also seems to be quite important having a score of more than 60 out of 100.***



Conclusion: The confusion matrix on the holdout data set shows the predictive power of the logistic regression model to be quite good and robust. With kappa of about 92% we are getting an accuracy as high as 97.5%. So finally, the logistic model should be productionalised as it is a very simple model with an excellent accuracy.