**Final Analysis Report for AGCO Case Study**

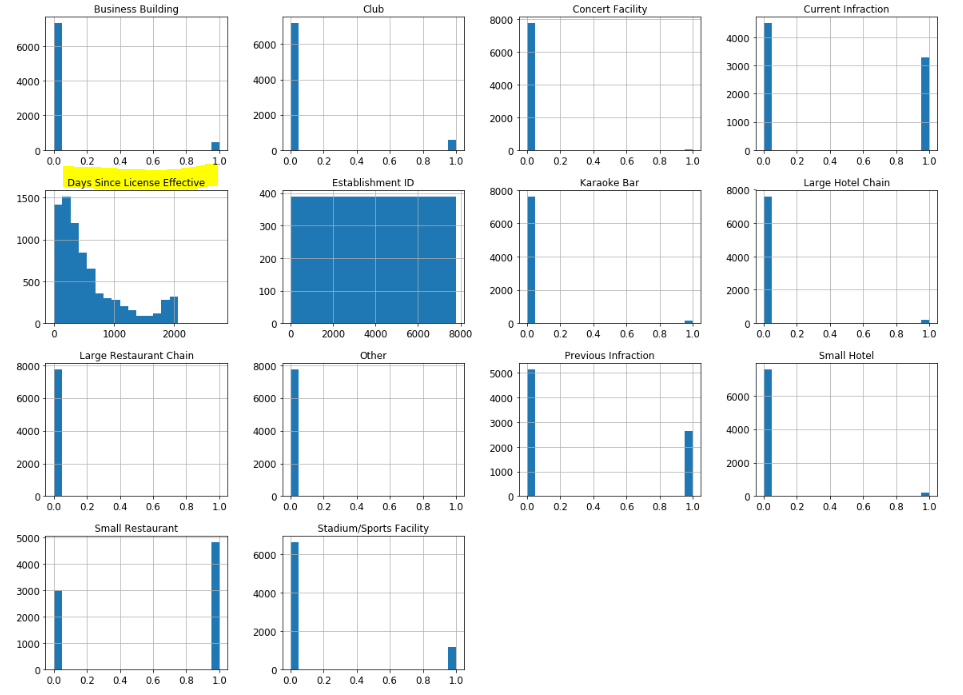
**By Ashish Gupta**

**Business objective**

The business objective of this assignment is to produce significant insights that will shed light on establishments’ infractions. In the given dataset, the Y variable or the Response variable is “**Current Infraction**”. For my analysis, a few variables have been removed. These are Establishment ID, Email on file and Postal First. The rest of the variables are “X” or Explanatory variables

**Exploratory Analysis of Data**

Before building the model, I decided to visualize the data to check for anomalies in data to detect outliers, influencers or even the skewness in. the variables. I found that the variable, “The number of days since the license is effective” is not normally distributed.

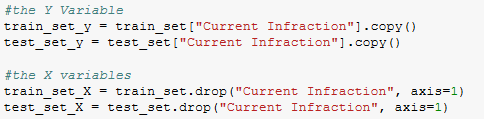


So it is important to transform this variable before building the model. A transformation may be used to reduce skewness. A distribution that is symmetric or nearly so is often easier to handle and interpret than a skewed distribution.

**Data Split: Training and Test Data**

Since I have more than 7000 records, I preferred creating a stratified sample based on the “Y” variable i.e. “Current Infraction”

I further created a train – test split using 80:20 scenario and further X and Y variables were separated to create the 4 separate data frames. (Two explanatory and two response data frames)



**Apply Transformation to “The number of days since the license is effective”**

Using the standard scaler, I transformed the values – while using the Pipeline and a Data frame Selector, that treat the transformation on selected variable only, keeping the other variables untransformed. Using the Feature Union, I merged the transformed variable to the existing explanatory training data frame.

**Build the Models**

Once my data processing was complete, I started building Logistic Regression model. Logistic Regression is easier to interpret and can generate a list of significant variables which drives the model and produce meaningful insights. Using the backward selection technique, I was able to derive the significant variables on the model.

|  |  |
| --- | --- |
|  | Odds Ratio    Confidence Interval |

**Interpretation and Insights of Logistic Regression**

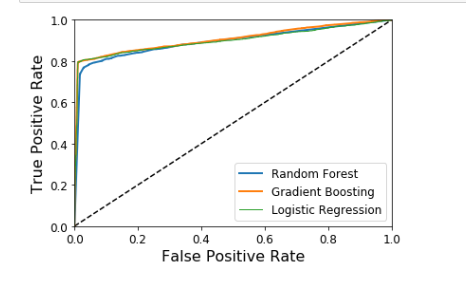
From logistic regression, I identified that establishments like Concert Facility and Large restaurants are not significant to the model, while variables like Club, Business Building, Small Hotel, Small Restaurant, Large Hotel Chain, Concert Facility and Stadium/Sports Facility play a significant role in determining the infractions. The number of days since the license is effective also plays a significant role in determining if infraction would occur or not.

**Improvement over Logistic Regression: Gradient Boosting and Random Forest Models.**

Taking the Logistic Regression model as my base model, I created Gradient Boosting and Random Forest Models. These additional models will be useful in accessing the best model of the three models. The measure for assessing the best models is roc\_auc. The outputs of model comparison will be highlighted later. Before that I had to tune our existing models for better performance.

**Tune the models using Grid search.**

I used different hyper parameters to tune the three models in order to get the best estimators that improve the existing model performance. Other than the hyper parameters, Grid search also uses cross validation folds as 4 to get a better fit on the training set. The models are further evaluated on the training set using cross\_val\_predict function and I further generate the probabilities of the positive class – which will be used in generating the roc curves.



AUC for the random forest classifier is .894612

AUC for Gradient Boosting is .9063312

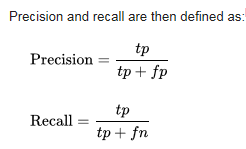
AUC for Logistic Regression is .898645

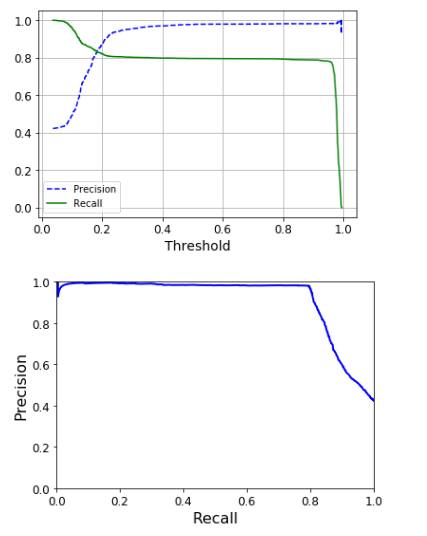
Based on the AUC score, I can say that Gradient Boosting model is a better model over Random Forest and Logistic Regression.

**Confusion Matrix and Precision/ Recall Curve**



From the confusion matrix above, I can see a good number of True positives (3556) and false negatives (2089) correctly predicted on the training set, I also found the Precision to be .9775 and recall to be .7958.





As per our business problem, we want our model to predict more Infractions to be predicted. As per the formula – the best way to increase our recall would be decrease false negative and improve true positives. So I changed my threshold values to increase my recall to .939047. Further this new change in the threshold is incorporated to the model by predicting the training set and test set on Gradient Boosting model. The accuracy on the training and test set is 90%

**Final Recommendations**

1. Since the number of days since the license is effective is significant, the management should take proactive actions to monitor the establishment’s license duration before infraction could occur.
2. Previous infraction is significant and Management needs to monitor these establishments more proactively, who have a previous history of infractions.