Final Project Report

Group E8

05/04/2022

Introduction

Analysis

Before the real analysis on the dataset, the exploratory data analysis is necessary. In this step, missing values are checked and removed. The dataset is cleaned to make sure it is well-structured for fitting the model. We split the `RatingDate` variable into `Year`, `Month`, and `Day`. We also created the response vaeriable `Pass` contains values 0 and 1, where 0 represents the rating below 3, and 1 represents the rating value equal and larger than 3. After wrangling the dataset, several plots of exploratory and response variables are created. Since the response variable `Pass` is categorical, we agree that bar plots and histograms are more suitable to explore the data. The plots below are one histogram of `LocalAuthorityCode`, and two bar plots of `BusinessType` and `Pass`.

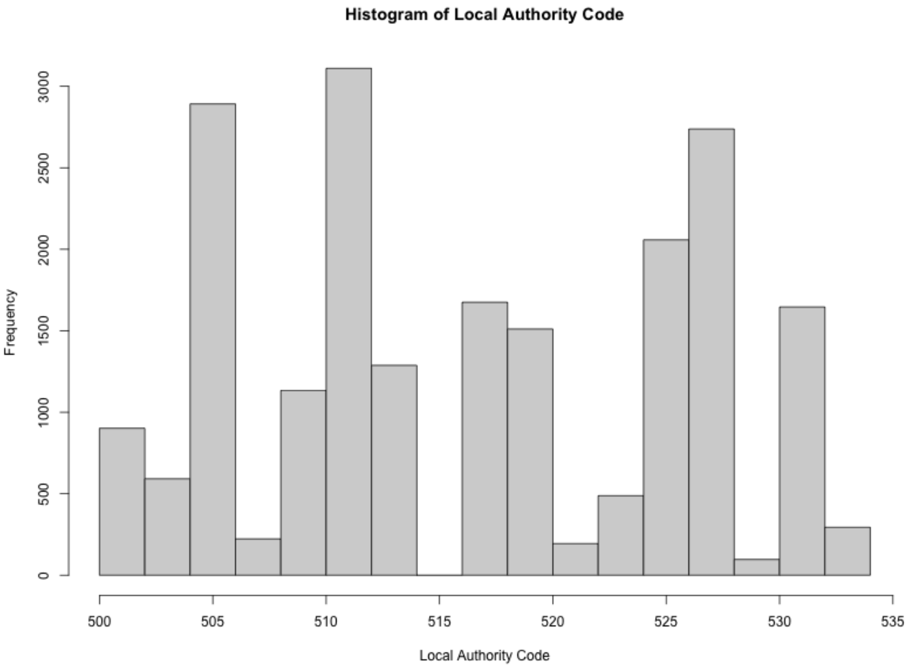

Figure 1: Histogram of Local Authority Code

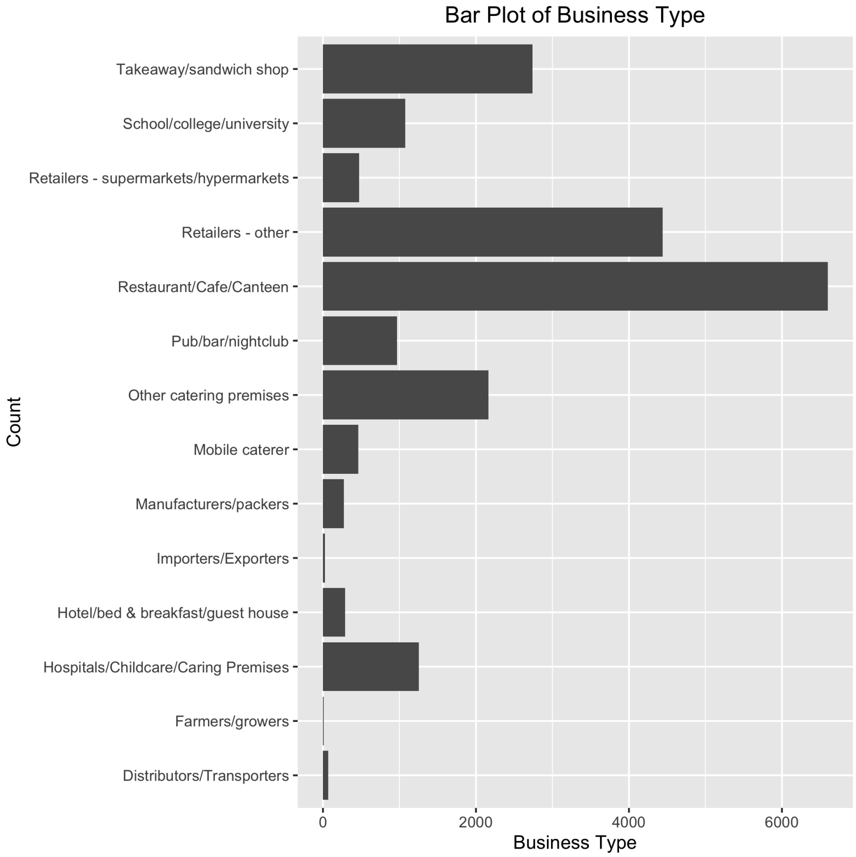

Figure 2: Bar Plot of Business Type

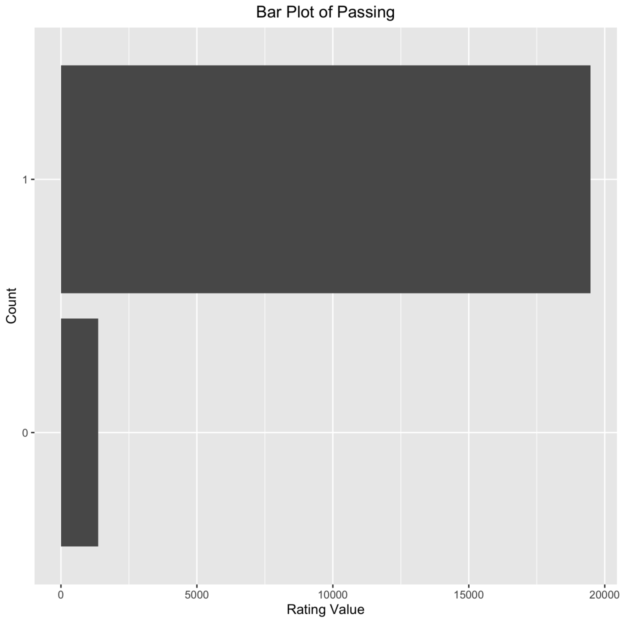

Figure 3: Bar Plot of Rating Value

We use train-test method to evaluate the final model in this project, hence it is essential to split the dataset into training and testing sets. We use 75% of the data to fit the model, and 25% of the data to test the prediction accuracy.

Since there are not many variables, we decided to train a full logistic regression model contains all the variables. There are only one variable that can be treated as a fector, which is the `BusinessType`. Although the `LocalAuthorityCode` are discrete numbers, and it seems be possible to be treated as a factor as well, the values variate and it will increase the complexity of the model. Hence, it is better to keep it as numerical variable. Moreover, it is no interaction terms in the model, and it is not intuitive to explore any pairs of interactions among these variables. After fitting the model, a summary of the coefficients, test statistics, and Akaike Information Criterion (AIC) values are obtained in the summary of the fitted model. The model summary is attached on the Appendix part.

The prediction results of the testing set are generated using the fitted model to evaluate the accuracy. A confusion matrix is also generated to show the number of true predicted and false predicted values. The matrix and the table of accuracy, as well as its interval is shown below.

```
##        Criteria         Value
## 1      Accuracy  0.9324635457
## 2         Kappa -0.0003827862
## 3 AccuracyLower  0.9253093094
## 4 AccuracyUpper  0.9391287058
## 5  AccuracyNull  0.9998081351
```

```
##   Prediction True_Value_0 True_Value_1
## 1          0            0          351
## 2          1            1         4860
```

The final accuracy of the model is 0.9324635, and the AIC is almost 7000.

Conclusion

Reference

Appendix - R Script

```r
# R code for the final project report
library(tidyverse)
library(ggplot2)
library(GGally)
library(rsample)
library(InformationValue)
library(caret)
# Read the dataset
data <- read_csv("./data/food_hygiene_rating_data.csv", col_names=TRUE)

# Clean the dataset
data_eda <- data %>%
  select(LocalAuthorityCode, RatingDate, BusinessType, RatingValue) %>%
  filter(RatingValue==0 | RatingValue==1 | RatingValue==2 |
           RatingValue==3 | RatingValue==4 | RatingValue==5) %>%
  mutate(Pass = ifelse((RatingValue >= 3), 1, 0)) %>%
  mutate(Year = format(RatingDate, "%Y") %>% as.numeric(),
         Month = format(RatingDate, "%m") %>% as.numeric(),
         Day = format(RatingDate, "%d") %>% as.numeric(),
         BusinessType = BusinessType <- as.factor(BusinessType)) %>%
  select(-RatingDate) %>%
  drop_na()
```

```
# Visualization of the dataset
png("results/LAC_hist.png",width=800, height=600)
LAC_hist <- hist(data_eda$LocalAuthorityCode,
                 xlab="Local Authority Code",
                 main="Histogram of Local Authority Code")
dev.off()

options(repr.plot.width = 15, repr.plot.height = 8)
ggplot(data=data_eda, aes(x=factor(BusinessType))) +
  geom_bar(stat="count") +
  xlab("Count") +
  ylab("Business Type")+
  ggtitle("Bar Plot of Business Type")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
ggsave("./results/BT_bar.png")

ggplot(data=data_eda, aes(x=factor(Pass)), height=3) +
  geom_bar(stat="count") +
  xlab("Count") +
  ylab("Rating Value")+
  ggtitle("Bar Plot of Passing")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
ggsave("./results/RV_bar.png")
```

```
# Split the dataset into training and testing sets
data_split <- initial_split(data_eda, prop = 0.75, strata = RatingValue)
training_data <- training(data_split)
testing_data <- testing(data_split)
test_X <- testing_data %>%
    select(-Pass)
true_vals <- testing_data$Pass

# Fitting the models
mod <- glm(Pass~LocalAuthorityCode+BusinessType+Year+Month+Day,
           data=training_data, family="binomial")

# Model evaluation
mod_sum <- summary(mod)
pred <- predict(mod, newdata=test_X, type = "response")
cut_off <- optimalCutoff(true_vals, pred)
final_predict <- ifelse(pred > cut_off, 1, 0)
final_predict <- as.factor(final_predict)
true_vals <- as.factor(true_vals)
conf <- confusionMatrix(true_vals, final_predict)
conf_mat <- as.table(conf)
acc <- as.matrix(as.matrix(conf,what="overall")[c(1,2,3,4,5),])
write.csv(acc,file="./results/accuracy.csv")
write.csv(conf_mat, file="./results/conf_mat.csv")
```