Asssignment 1

Nigel

2022-09-04

## Data

### 1. Describe your data visually and numerically

Both data sets contained mixed variables.

The weather data set consists of 422 observations and 22 variables. This data set records daily weather conditions including temperature, wind, humidity, rainfall, etc.

dim(weather\_data)

## [1] 422 23

The sales data set consists of 414 observations and 9 variables. This data set records daily sales and details the source of sales including UberEats, Menulog, Deliveroo sales, and cash and eftpos sales.

dim(sales\_data)

## [1] 326 10

### 2. Provide details of the pre-processing undertaken in order to meet the conditions of your chosen modelling techniques.

The variable names given to the weather data were renamed to aid in pre-processing.

The weather data set contained very few NA values.

## date min\_temp max\_temp rainfall evap sun\_sh\_hr   
## 0 0 1 2 0 1   
## dir\_max\_wind spd\_max\_wind tm\_max\_wind 9am\_temp 9am\_hum 9am\_cld\_amnt   
## 0 2 0 0 1 1   
## 9am\_wind\_dir 9am\_wind\_spd 9am\_pres 3pm\_temp 3pm\_hum 3pm\_cld\_amnt   
## 0 0 0 0 0 0   
## 3pm\_wind\_dir 3pm\_wind\_spd 3pm\_pres rain\_today rain\_tmrw   
## 0 0 0 0 1

We decided to resolve these NA values by imputing the monthly means for each year. To do this we created a “month” and “year” variable by extracting the values from the “date” variable.

The eight most recent observations were removed as the sales data set only records data eight days prior to the weather data set.

Two separate .csv files were used to create the sales data set. One recorded sales data for 2021 and the other for 2022. The 2022 file contained an extra variable recording eftpos sales including the surcharge. This variable was the “Card” variable multiplied by the surcharge (1.011%). This variable was removed as it was deemed unnecessary.

Next, the variable names were renamed to keep consistency during pre-processing.

We also created a “month” and “year” variable for this data set using the same method as the weather data set.

The sales data contained numerous NA values.

## date day ubereats menulog deliveroo card cash   
## 0 0 3 58 189 1 48   
## total petty\_cash month year   
## 0 50 0 0

The majority of these NAs occurred in the “uber”, “menulog”, and “deliveroo” variables. These NAs exist due to the restaurant being closed during the COVID-19 lockdown so it makes sense to replace them with 0.

For both data sets, a “yr\_month” variable - in the format “Jan 2000” - was created to allow the creation of time-series graphs.

Next we combine the two datasets so that we can create a correlation matrix to determine the significant variables for our chosen models. The “uber”, “menulog”, “deliveroo”, and “total” variables were appended to the weather data to create a master data frame. Next, the “rain\_today” and “rain\_tmrw” variables were modified to contain binary values. Then we removed the observations where total sales equals zero. Finally, we created a subset of the master data frame containing only numeric variables to allow for the creation of the correlation matrix.

## Warning in ggcorr(corr\_master\_df, label = TRUE): data in column(s) 'month',  
## 'yr\_month', 'date', 'month.1', 'yr\_month.1' are not numeric and were ignored



## Analyse

### 1. Describe the modelling procedure with enough detail to allow someone else to repeat the modelling

##   
## Call:  
## lm(formula = rainfall ~ min\_temp + max\_temp + sun\_sh\_hr + `9am\_hum` +   
## `9am\_pres` + `3pm\_temp` + `3pm\_hum`, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.950 -2.448 -0.941 0.483 39.354   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 143.16818 61.69182 2.321 0.0212 \*  
## min\_temp 0.28405 0.14417 1.970 0.0500 \*  
## max\_temp -0.57203 0.32375 -1.767 0.0785 .  
## sun\_sh\_hr 0.29114 0.15744 1.849 0.0657 .  
## `9am\_hum` 0.07435 0.02948 2.522 0.0123 \*  
## `9am\_pres` -0.14975 0.06089 -2.460 0.0146 \*  
## `3pm\_temp` 0.41099 0.34832 1.180 0.2392   
## `3pm\_hum` 0.07724 0.04061 1.902 0.0584 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.804 on 235 degrees of freedom  
## Multiple R-squared: 0.1488, Adjusted R-squared: 0.1235   
## F-statistic: 5.87 on 7 and 235 DF, p-value: 2.696e-06

## [1] 2.5266

We chose two models for our prediction which are linear regression and decision tree. As we can see that rmse in model 1 is 2.5266, p value is 1.686e-06 which has shown that our features are useful for our prediction in the next part.

##   
## Call:  
## lm(formula = rainfall ~ min\_temp + max\_temp + sun\_sh\_hr + `9am\_hum` +   
## `9am\_pres` + `3pm\_temp` + `3pm\_hum`, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.950 -2.448 -0.941 0.483 39.354   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 143.16818 61.69182 2.321 0.0212 \*  
## min\_temp 0.28405 0.14417 1.970 0.0500 \*  
## max\_temp -0.57203 0.32375 -1.767 0.0785 .  
## sun\_sh\_hr 0.29114 0.15744 1.849 0.0657 .  
## `9am\_hum` 0.07435 0.02948 2.522 0.0123 \*  
## `9am\_pres` -0.14975 0.06089 -2.460 0.0146 \*  
## `3pm\_temp` 0.41099 0.34832 1.180 0.2392   
## `3pm\_hum` 0.07724 0.04061 1.902 0.0584 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.804 on 235 degrees of freedom  
## Multiple R-squared: 0.1488, Adjusted R-squared: 0.1235   
## F-statistic: 5.87 on 7 and 235 DF, p-value: 2.696e-06

## [1] 3.024834

rmse in model2 is 3.024 which is almost the same as model1 which we can use both model for our prediction in the next part. 