**Final project Group 25 Meeting minutes**

**Team members** Xinyi JIN (Melody) Yan YING (Eliza) Jiayuan ZHU (Anna)

1) 09/09/2020 (Wednesday) 10pm – 12pm

Agreed on general preference of the dataset,

e.g. well-purposed regression task; high-dimensional dataset with missing values;

Selected and discussed candidate datasets for the project.

Second used car (Kaggle)

Beijing Air quality (UCI)

Earthquake prediction (Kaggle)

DonorsChoose (Kaggle)

Great energy predictor (Kaggle)

(Melody)

2) 10/09/2020 (Thursday) 11am-1pm

Agreed on the potential selection for dataset, the earthquake dataset

Explored basic information of earthquake dataset and found external library which contained additional earthquake information. Further exploration of external library installation and functionalities.

Discussed teamwork preference:

Meeting schedule: EVERY Wed, Fri, Sun at 10pm

Github repository set up

(Melody)

3) 11/09/2020 (Friday) 10pm - 1am

Discussed limitations and other considerations of the earthquake dataset

Found an alternative dataset related to alcohol selling in Iowa and simply analysed the advantages of its attributes

Confirmed this one as the appropriate dataset for final project

Came up with several potential visualization and prediction tasks and started proposal writing

Planned to continuing proposal discussion on Sunday

(Anna)

4) 13/09/2020 (Sunday) 10pm - 12pm

Confirmed the structure of the proposal for the project

Discussed the specific direction that will be focussed on for future analyzation

Justified the details of the identify tasks and method

Confirmed the identify task to be identifying most profitable liquor brands and associated volumes to sell in Iowa based on location and time

Discussed whether Lasso method is suitable for selecting features in brand

Planned to complete the proposal before Wednesday (deadline is at 5p.m Friday)

Planned to check and submit proposal on Wednesday meeting and start data processing on the dataset

(Eliza)

5) 17/09/2020 (Thursday) 11am-2pm

Discussed potential problems of methods for previous proposal and modified the specific tasks.

Further discussed details for the updated proposal and finalised the proposal.

Re-selected the subset used for the project

Assigned potential tasks to individuals

(Melody)

6) 20/09/2020 (Sunday) 10pm-12pm

Discussed the real-life meanings of the attributes and understanded potential relationship between them

Loaded the dataset and checked data types

Found missing values then removed them

Found three counties with most selling records with groupby method

Convert ‘Date’ attribute to several variables (year, month, day)

Made county number consistent with county name (inconsistency data may be caused by typo)

Convert all attributes to the desired formats

Removed potential error data (bottle volume, bottle cost, bottle retail, sales, bottle sold, volume sold being zero)

(Anna)

7) 23/09/2020 (Wednesday) 10pm-12pm

Continued the data pre-processing

Tried to use the index method to separate the latitude and longitude from the “store location”

Met some difficulties so decided to continue work in the lab on Thursday

(Eliza)

8) 24/09/2020 (Thursday) 11am-2pm

Changed the method to be regular expression to choose latitude and longitude and store them to be new columns in the data frame

Downloaded the shape file of Iowa for choropleth maps

Followed the proposal written, plot the choropleths of total alcohol sales in different countries and total alcohol profits in different countries

Planned to finish the overall graphs of the filtered data frame next time and arrange individual work

(Eliza)

9) 25/09/2020 (Friday) 10pm-12pm

Continued the data cleansing process.

Noticed mismatches between store number and store name, had some problems fixing this mismatch (plan to fix it tomorrow)

Discussed potential limitations and problems with the previously proposed models

Redefined the response variable to predict, and discussed the candidate models to implement.

Re-scheduled the meeting plans afterwards (new meeting plan: Every Thursday(11-2pm)/ Saturday(5-7pm)/ Sunday (5-7pm))

(Melody)

10) 26/09/2020 (Saturday) 5pm-7pm

Changed the proposed models and clarified the research task

Modified the pre-processing steps and found more mismatches between store number and store name

Followed the same procedure as before but unexpected errors occur. Tried to fix errors, but found that some store number may match different store names (guess: typo for store names) or some store names may have different store numbers (some stores may share the same name)

Decided to utilize store numbers and referred to store names in the analysis stage

Made the attribute correlation plot

Rethought the appropriate graphs to draw

(Anna)

11) 01/10/2020 (Thursday) 11am-2pm

Discussed the requirement of the model description and the summary statistics of attributes

Discussed the overall structure of the report for the final project as well as the following process of the code

Asked the tutor some related questions about the final project like the purpose of comparing the models and the ideas needed to be explained

Planned to finish the overall graphs and start on the models

Planned to start the pre-processing and cleaning part in the report as well

(Eliza)

12) 03/10/2020 (Saturday) 10pm-1am

Further cleansed the dataset, with particular attention for outlier detections.

Investigated descriptive statistics of each numeric attributes and selected four suspicious columns (volume sold (response)/pack/retail price/ cost price) for interquartile range calculation (data records beyond the 3IQR were discarded); Discussed physical reasons for some extreme values for attribute “bottle volume”, and agreed to keep the extreme values

Correlation plots using numeric attributes, noticed some potential transformations for both response and predictors

Built an initial linear regression model with all features. Discussed the adjusted R square and p-values of each attribute.

Noticed latitude and longitude do not have significant p-values and decided to explore further next meeting.

(Melody)

13) 04/10/2020 (Sunday) 5pm-8pm

Created a new attribute relates to the number of stores in each county

Tried to create another predictor variable corresponds to the total volume sold in each county, but realized that it would be based on the response variable and then forgo this attribute

Continued outlier removal. Searched the boundary values for Iowa State and then remove longitude and latitude outliers according to this standard

Generated the correlation plot with numeric attributes and the output was better than the previous plot

Computed the baseline model and the corresponding RMSE

Constructed several base models with all features and calculated the RMSE. The models include Linear Regression, Random Tree and Random Forest from Scikit learn library

Decided to choose the final model for this project next time

(Anna)

14) 08/10/2020 (Thursday) 10pm-12pm

Discussed the assumptions about different regression models, especially linear regression. Since the linear regression requires the data to be with a linear relationship, unbiased error and others, linear regression may not be so good for our data.

Used cross validation to evaluate the performance of different parameters in Decision Tree Regressor.

Plotted line charts of root of mean squared error vs. values of parameters to find the appropriate range of better parameters in the models to have smaller RMSE in the line graph.

Computed parameter selection tool - “Bayesian Optimization” on selected ranges of parameters to find out the best combination of parameter values in Decision Tree Regressor which gives the smallest root of mean squared error

The parameter tuning will be processed in the same steps for different models including plotting line charts and using parameter selection tool.

Planned to set the number of cross validation folds to be 3 for selected models like Decision Tree, Random Forest, and others next time.

(Eliza)

15) 10/10/2020 (Saturday) 10pm-12pm

Further applied Bayesian Optimization to the decision tree regressor, completed the RMSE versus regularization plots for two major parameters for decision tree regressor (maximum depth and maximum feature percentage).

Have not finalized the tuning process (Bayesian optimization is heavily time consuming).

Planned to apply Bayesian optimization techniques to random forest regressor as well.

Assign different report parts to each team members and discussed rough timeline for the completion of each task.

Discussed potential coverage and scope for each part of the report, and the total page number restrictions

Recorded several questions that may require specific answer from the tutor.

(Melody)

16) 11/10/2020 (Sunday) 5:30pm-6:30pm

Created the dataset after pre-processing and randomly sampled one million records

Calculated the correlation between response variable and category attributes

Confirmed the final models to implement (Random Tree, Random Forest, LightGBM)

Started report planning and assigned each team members’ tasks

Abandoned Google Drive and transformed all materials to Teams

Decided to finish coding at the end of Thursday and started PPT making

Planned to finish PPT organization at the end of next Monday and practice video presentation

(Anna)

17) 15/10/2020 (Thursday) 11am-2pm

Found the problem of the distribution of sold volume in 2017 through boxplots

Printed out the month, category, bottle volume and vendor information from 2014-2017 to find any related reasons about the distribution problem of 2017

Found that 2017 data does not contain records from November and December

Also found that categories in 2017 produced bottles with much smaller volume. As a result, the average volume sold in 2017 becomes much smaller than that of previous years. The reason behind that may be large-volume bottles are not so popular or shops changed strategies.

Asked tutors about some specific questions about recording and report, especially on the performance or format requirement of these two things.

Discussed the plan to use plots in the report including whether to put them into appendix and the number of graphs to put

According to the arranged tasks, each team member does own part separately

Planned to discuss the work have been finished and start to make slides and plan of recording next time

(Eliza)