ML VAPING EFFECT PROJECT WEEK 9 REPORT

Prepared By: Ng Jun Kiat (u7338876)

OVERVIEW

1. This report describes (1) the models built, (2) the model building process, (3) results of model testing, (4) the importance of each input feature, and (5) potential follow-up actions.

MODELS

2. Three models were built and tested: (1) Random Forest Regressor without bootstrapping, (2) Random Forest Regressor with bootstrapping, and (3) Extreme Gradient Boosting (XGBoost). When bootstrapping is not used, the entire dataset is used to train each decision tree. When bootstrapping is used, only a subset of the dataset is used to train each decision tree. These samples may be used again for the training of another decision tree. In XGBoost, decision trees are generated iteratively. Each decision tree will predict the output value, and the residuals will be considered by the next decision tree when making a prediction. This way, each newly generated decision tree will be better than the previous decision tree.

METRIC

3. The metric used is Mean Absolute Percentage Error (MAPE). The formula is shown below. A value of zero indicates that the actual values and the predicted values are the same for all samples. The larger the value, the larger the error between the actual values and the predicted values. For instance, a MAPE of 0.5 would indicate an error that is 50% the size of the actual value. A MAPE of 0.1 or less would be acceptable, while any MAPE above 0.5 is too high.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

M = mean absolute percentage error
n = number of times the summation iteration happens
A_t = actual value
F_t = forecast value

Commented [JN1]: I've verified that MAPE is the correct term for the formula:

https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

PROCESS

4. Before training the model, 5-fold cross validation is used to find the best hyperparameters. The hyperparameters that were finetuned include: (1) *n_estimators* which indicates the number of decision trees used in the respective models, and (2) *max_depth* which indicates the maximum depth of each tree. In cross-validation, the dataset is partitioned into 5 sections, each section taking turns being the validation set. When a particular section is used as the validation set, the remaining sections are used as the training set. The MAPE calculated is the average of the validation score of all five validation sets. After tuning the hyperparameters, the final model is trained with 80% of the dataset, and tested with 20% of the dataset.

RESULTS

5. *N* refers to the number of times the dataset was duplicated before adding Gaussian noise. **Figure 1** shows the (MAPE) of the model in Quality Adjusted Life Years (QALYs) and **Figure 2** shows the MAPE of the model in Health System Cost Savings.

Figure 1: MAPE of Models in QALYs

Dataset Duplicates	N = 1	N = 3	N = 5	N = 10
Random Forest without Bootstrapping	**MAPE=2.33** n_estimators=60 max_depth=10	**MAPE=0.47** n_estimators=10 max_depth=20	**MAPE=0.11** n_estimators=30 max_depth=15	**MAPE=0.03** n_estimators=80 max_depth=20
Random Forest with Bootstrapping	**MAPE=2.10** n_estimators=90 max_depth=20	**MAPE=4.21** n_estimators=10 max_depth=10	**MAPE=0.89** n_estimators=30 max_depth=15	**MAPE=0.42** n_estimators=20 max_depth=20
XGBoost	**MAPE=2.53** n_estimators=70 max_depth=15	**MAPE=6.07** n_estimators=40 max_depth=10	**MAPE=1.16** n_estimators=90 max_depth=5	**MAPE=0.17** n_estimators=40 max_depth=10

Figure 2: MAPE of Models in Health System Cost Savings

Dataset Duplicates	N = 1	N = 3	N = 5	N = 10
Random Forest without Bootstrapping	**MAPE=15.9** n_estimators=10 max_depth=10	**MAPE=10.7** n_estimators=50 max_depth=10	**MAPE=1.29** n_estimators=10 max_depth=20	**MAPE=1.22** n_estimators=10 max_depth=10
Random Forest with Bootstrapping	**MAPE=7.06** n_estimators=10 max_depth=5	**MAPE=9.69** n_estimators=40 max_depth=20	**MAPE=3.16** n_estimators=10 max_depth=20	**MAPE=3.42** n_estimators=10 max_depth=10
XGBoost	**MAPE=7.42** n_estimators=30 max_depth=5	**MAPE=11.5** n_estimators=20 max_depth=10	**MAPE=2.07** n_estimators=70 max_depth=15	**MAPE=1.62** n_estimators=40 max_depth=15

DISCUSSION

6. To recap, the model in the Week 8 Report had a QALY MAPE of 0.20 and a Health System Cost MAPE of 1.53. The chosen model should ideally have a better performance. It appears that Random Forest without Bootstrapping is the preferred model. However, we must note that in this method, all training data is used in every decision tree. Given that the dataset is duplicated multiple times, it is likely that every sample in the original dataset is used to train the model, making it likely to overfit. On the other hand, each decision tree in XGBoost uses only half of the training set, making it more robust. In addition, the XGBoost model still gives a MAPE of 0.17, which is positive. One possible action item could be to further optimise the XGBoost model. Currently, only the number of decision trees and the maximum depth of those trees are optimised. More hyperparameters include the normalisation to use, step size, minimum size of lead node, etc.

FEATURE IMPORTANCE

7. In both predictors of QALYs and Health System Cost, *dec_smoking_prevalence*, which indicates how much the policy results in a decrease in smoking prevalence among the population, is the most important feature. What is worrying is the insignificance of the *dec_smoking_uptake* parameter, which is an important feature to parameterise vaping policies. One possible action could be to remove the *dec_smoking_prevalence* feature altogether and retrain the model, or to

artificially increase the importance of the *dec_smoking_uptake* parameter. **Figure 3** shows the feature importance for the QALY model and **Feature 4** shows the feature importance for the Health System Cost model.

Figure 3: Feature Importance for QALY Model

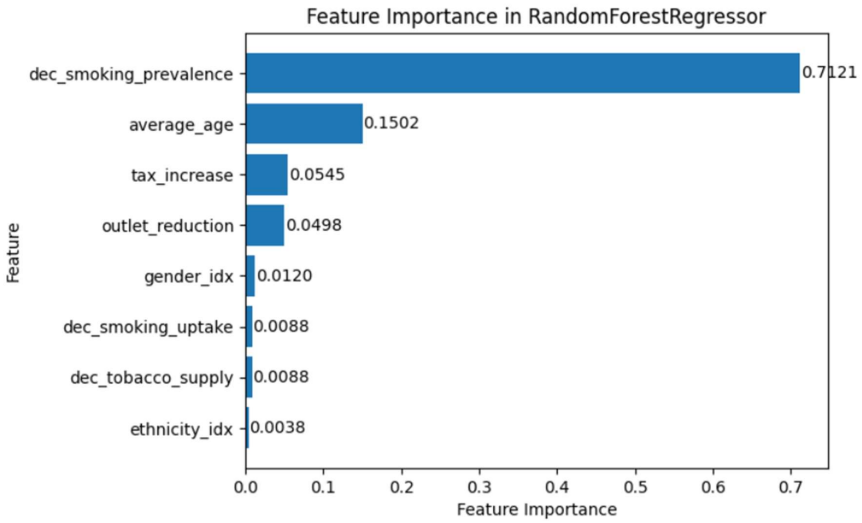
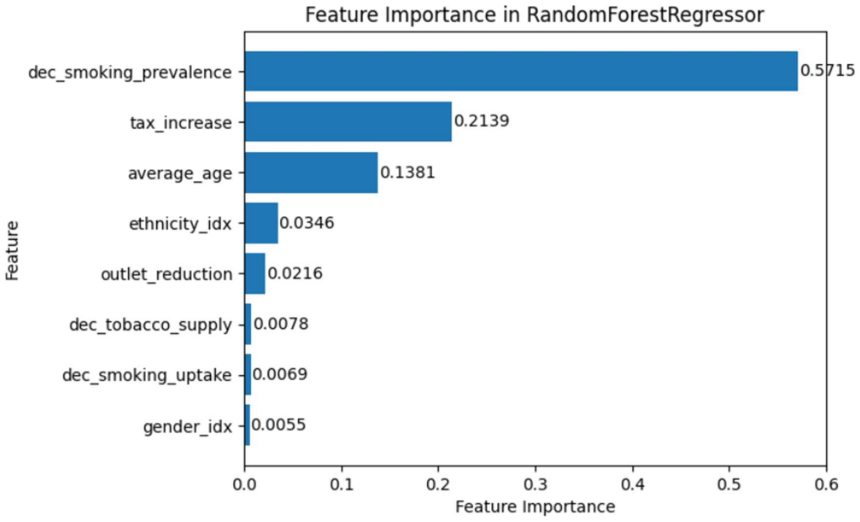


Figure 4: Feature Importance for Health System Cost Model



FOLLOW-UP ACTIONS

8. My presentation will be on 29 Oct (Tuesday), which is one week after Week 12. Here is the proposed timeline from now till the presentation:

- Week 10 Report: Further optimise XGBoost
- Week 11 Report: Produce Initial Findings using Vaping Research Paper
- Week 12 Report: Update on Presentation Progress