ACSC/STAT 3740, Predictive Analytics

WINTER 2024

Toby Kenney

Homework Sheet 3

Due: Wednesday 6th March: 11:30

Note: This homework assignment is only valid for WINTER 2024. If you find this homework in a different term, please contact me to find the correct homework sheet.

Standard Questions

1. An online shopping company is building a recommendation system to suggest purchases. It has collected the following data in the file `HW3Q1.txt`.

Variable	Meaning
category	The category of the item
market	The buyers targetted by the product.
brand.quality	The brand quality of the item: 0=no branding, 1=top quality
popularity	A measure of the number of users who purchase the item.
user.ID	A unique identifier for each user in the system.
month	The month of the year.
recent.spending	The amount spent by the user in the past month.
price	The price of the product.
purchase	Whether the user purchases the item.

   (a) Fit a random forest to predict whether the user will purchase the item. Use this to predict the probability that the user will purchase the item for the cases in the file `HW3Q1_test.txt`.

   (b) A natural feature to add to this data set is the ratio `price/recent.spending`. Refit the random forest predictor with this feature added, and use this to estimate the probabilities for the test data.

2. The file `HW3Q2.txt` contains data from a study on the corrosion of alloys under various conditions. The variables included are

Variable	Meaning
iron.percent	The percentage of iron in the sample.
chrome.percent	The percentage of chrome in the sample.
temperature	The temperature at which the sample is kept.
humidity	The humidity at which the sample is kept.
salt.concentration	The concentration of salt in the air around the sample.
light.intensity	The intensity of light to which the sample is exposed.
sample.impurity	The percentage of impurities in the sample.
one.hour.oxidation	The percentage of the sample which oxidises in a 1-hour period.
ten.hour.oxidation	The percentage of the sample which oxidises in a 10-hour period.

Fit a generalised linear model to predict whether the one hour oxidation and ten hour oxidation will be non-zero, and conditional on these being non-zero, fit a GLM with a gamma distribution for the conditional distribution of the one-hour and ten-hour oxidisation. Use this to predict the oxidation for the data in `HW3Q2_test.txt`.

3. The file `HW3Q3.txt` contains daily maximum temperature recordings in a certain city.

   (a) Fit a seasonal trend using the function $\sin(2\pi t)$ and $\cos(2\pi t)$ where t is the time in years, and a linear trend to reflect global warming.

   (b) After subtracting the seasonal and linear trends, fit an ARMA model to the residuals, using AIC to determine the best choices for p and q.

   (c) Fit a GARCH model to model the variance.

   (d) Based on this model, what is the probability that the average temperature in July 2026 will exceed $30°C$? [You can use the `ugarchboot` function to run a simulation to estimate this.]

4. A provincial government has collected the following data on the effect of regulations on economic activity in the file `HW3Q4.txt`.

Variable	Meaning
interest.rates	The prime interest rates
unemployment	The unemployment rate
consumer.confidence	An index measuring consumer confidence
CPI	Annual consumer price inflation over the previous 12 months
stock.index.returns	The increase in the stock market index over the past 12 months
safety.regulations	An index measuring the number of safety regulations in place for businesses
other.regulations	An index measuring the number of non-safety regulations in place for businesse
GDP	The GDP growth during the following 12 months

   Fit a generalised additive model to predict the GDP growth, using a normal response variable and identity link function.

   Use this model to predict GDP growth for the cases in the file `HW3Q4_test.txt`.

5. A life insurance company has collected the following data on the effect of particulate matter pollution on mortality. The data are in the file `HW3Q5.txt`.

Variable	Meaning
year	The year of the data
age.group	The age range of the population in question
location	The city being studied.
high.temp	The highest temperature during the year
low.temp	The lowest temperature during the year
particulate.matter	The average amount of particulate matter in the air
mortality	The percentage of this age group who died during the year.

(a) Fit a decision tree to predict mortality from the other variables.

(b) Convert age group to numeric, add year of birth as a predictor, and log-transform the mortality rate, and refit a decision tree.

(c) Fit a random forest model to predict log mortality from the other variables, including year of birth. Use this model to predict mortality for all age groups in Toronto in 2025 if the high temperature is $37°C$, the low temperature is $-11°C$, and particulate matter is 133.