Exercise 05

Section 1: Logistic Regression

The dataset represents data from the Framingham Heart Study, Levy (1999) National Heart Lung and Blood Institute, Center for Bio-Medical Communication. Researchers are interested in studying risk factors for coronary heart disease (CHD).

Position	Variable	Variable Label	Codes
1.	id	Patient identifier	
2.	sex	Patient gender	1 = male 2 = female
3.	sbp	Systolic blood pressure, mm Hg	
4.	dbp	Diastolic blood pressure, mm Hg	
5.	scl	Serum cholesterol, mg/100 ml	
6.	age	Age at baseline exam, years	
7.	bmi	Body mass index, kg/m^2	
8.	month	Month of year of baseline exam	
9.	chdfate	Event of CHD at end of follow-up	1 = patient developed CHD at follow-up 0 = otherwise

1. Answer the following:
    a. What is the outcome?
    b. What are the predictors researchers are interested in?
    c. What is the hypothesis?
2. Import the data, print out a few rows, and compute summary statistics. Is there missing data or other concerns?
3. Month of the year at baseline is an unwieldy variable meant to adjust for seasonal effects. Rather than put it in the model as is, create 4 binary variables for each season. This link will give examples of how to do this. The categories should be winter, spring, summer, & fall and should be defined as follows based on the month:
    a. Winter: 12, 1, 2
    b. Spring: 3, 4, 5
    c. Summer: 6, 7, 8
    d. Fall: 9, 10, 11
4. Fit a logistic regression model using all the relevant predictor variables (Note: use season, not month. Also, ID is not a predictor variable. Do not use it). Use statsmodels for now.
5. Conduct model diagnostics. This reference may be helpful.
    a. Look at distributions of the main predictor variables (excluding the new season variables). Do any require transformation?
    b. Check to see if collinearity is present. Explain what you find.
    c. Check linearity for each of the continuous covariates. Do those covariates each have a linear relationship with the outcome?

    d. Are there outliers?
    e. Are there at least 5 outcomes per category of sex?
6. Fix any issues you find and refit the model.
7. Compute the ORs and their confidence intervals. Interpret the ORs.

We will be using the county_level_election.csv dataset. This is 2016 election data and we are going to measure 'votergap' as the outcome. 'votergap' = trump-clinton. The exercise will build on the work from the decision tree lab.

Section 2: Bagging / Random Forest

We are going to be using test and training splits, cross validation, and fitting a random forest to the data. Create an 80/20 Train/Test split. For accuracy use the .score method.

```
from sklearn.ensemble import RandomForestRegressor
```

1. Set the number of estimators to be 100, the features to be the square root of available features, and iterate through depths (1-20). Use only 5 folds for cross validation to save some compute resources. Plot the max depth on the x axis and the accuracy on the y axis for training and for the mean cross validation.
2. Based on the plot, how many nodes would you recommend as the max depth?
3. What is the accuracy (mean cv) at your chosen depth?
4. The cross validation looks different than the lab, why?

Section 3: Boosting / XGBoost

```
import xgboost as xgb
```

5. Use the defaults for most parameters. Iterate through depths (1-20). Use only 5 folds for cross validation to save some compute resources. Plot the max depth on the x axis and the accuracy on the y axis for training and for the mean cross validation.
6. Based on the plot, how many nodes would you recommend as the max depth?
7. What is the accuracy (mean cv) at your chosen depth?
8. The cross validation looks different than random forest, why?