Answers to Task 5.5: Intro to predictive analysis

By: Alexandra Lindsay

Step 1: Understanding Regression

Conduct some research on logistic regression and explain how it differs from linear regression. When would you use logistic instead of linear regression and why?

To recap linear regression is a relationship between a depended variable and one or more independent variables. In other words, it is to see how different variables affect the main one. The variables used must be numerical data and would be featured in a linear model. Most of the time it is to analyse current or past data to forecast the result. As an example, you could use this to determine the trends in the fluctuations of the exchange rates.

Logistic regression is used to calculate a binary probability like a yes or no question. For example, do our customers have a credit card: yes/no. the logistic regression is used to predict the two possible answers. Contrary to linear regression the independent variable(s) cannot have any correlation to the dependent variable. It falls under a classification model.

In the case of Pig E.E. bank, using the linear regression to find out the trends in the fluctuations of exchanges rates or using logistic regression to calculate if a customer has a credit card, is necessary due to the genre of the question but also the data. In the end they both help companies make informed decisions.

Step 2: More on Linear Regression

Based on the R2 score of 0.86 being this close to 1 the accuracy of the model is quite good. Also, the alert volume shows a weak positive relationship to the number of recorded customers. Although, there are only 10 recorded observations in this model, I would recommend including or recording more data to verify its accuracy.

Step 3: Differentiating between Models

- Scenario A: I would you a linear regression model. The global oil prices as the dependent variable and the independent could be the unemployment rate, the GDP in the top 20 countries and more…
- Scenario B: I would use a classification model and use a logistic regression to predict which movies would fit each user. It is essentially to build a customer profile base on previous movies or tv shows they have watched. So, your dependent variable would be if they liked this movie and as independent variables would be what they previously watched or even what other users that have the same tastes as them and what they liked….

Step 4: Bias in Your Data

Since the alert flags are not defined, they could lead to multiple investigator biases like: customer profiling and geographical location, job, salary, etc. We also don’t know where this population (of 10) sample was taken, for all we know it was in a highly influential or poor neighbourhood. Therefore, not representative of the general population and could have unmentioned exclusions.