---
title: "ASSIGNMENT 1"
author: "Bianca Oguh"
date: "2023-12-05"
output:
  pdf_document: default
  html_document: default

DS METHODOLOGY CHOICE
A CRISP-DM methodology is employed to report this study as it is comprehensive, and emphasises the importance of iteration in processes throughout data mining projects. While KDD is a more detailed methodology than CRISP-DM as it has 3 additional stages, KDD is primarily applied in projects that require extraction of previously unknown information, however the general scope of this project is defined clearly from the beginning (Shafique and Qaiser, 2014). SEMMA is considerably simplistic, as it fails to consider inclusion of data manipulation techniques for optimal results, failure of data understanding and processing could lead to inaccuracies in the model, and thus inaccuracies in predictions.

BUSINESS UNDERSTANDING
Social proof methods such as online product reviews builds trust within the buyer-seller relationship, and remains one of the most important factors consumers consider before making a purchase with a business. As internet availability increased over time, a growing number of consumers leave reviews about a business on platforms such as Yelp which published crowd-sourced reviews about a business. Predicting user reviews benefit both consumers and businesses: quantifying the choice of words in a review streamlines consumer's decision making when comparing the qualities of similar businesses, and businesses can acquire useful insight into approximating how different sentiments are put on a one to five scale. And so this project aims to use the texts of user reviews to predict the number of stars an individual i rates a business j.

MODELLING
Due to the discrete nature of the 'stars' variable, classification techniques are employed to produce consistent confusion matrices when calculating accuracy. Therefore a classification technique for review predictions is employed: more-specifically Multi-Class Logistic Regressions.

Multi-Class Logistic Regression
This technique is chosen due to its relative simplicity compared to constructing decision trees, and intuitive approach. This technique extends typical binary outcomes where there are 2 classes, where K=5 in this instance. The 5 classifiers are trained by differentiating between "their class" of which a value of +1 is assigned, and "not their class" of which a value of -1 is assigned.

EVALUATION
Multi-Class Logistic Regression
The model predicted ratings accurately 65% of the time. Assuming independence of document features to classifications, the Multi-Class Logistic Regression model can be a robust method to estimate user ratings based on probabilities. However, some caveats present while estimating this model. The distribution of one to five star reviews is heavily skewed towards 4 and 5 stars, thus making the prediction of these ratings far more accurate as shown by the confusion matrix relative to the other ratings. Also the independence condition does not account for negation handling cases, especially when tokenising documents to unigrams, impeding the effectiveness of the model. Thus the accuracy count may be improved using POS (part of speech) techniques, and higher n-grams.

DEPLOYMENT
The relative simplicity of this technique helps to make deployment techniques less of a challenge. But, the feasibility of running these codes is dependent on available memory, as while larger datasets provide closer to true representations of estimating how user reviews map to star rating, issues can arise in steps such as vectorisation of data if memory capacity is insufficient to do so. However, larger businesses can better afford advanced infrastructure to handle large sets of data.

CHALLENGES
While R being an open source programming language is beneficial in terms of finding solutions to specific problems in community forums such as Stack Overflow as there are

a relatively large number of contributors, a lack of clear understanding of specific objectives limits the ability to understand what components are required to achieve these objectives. Simple tasks such as merging datasets were immediately solved upon searching, but to find how to make a model interpret categorical numerical data on a multi-class scale, or how exactly to prepare my data to be able to produce accurate results was more difficult.

To overcome this, I conducted a literature search on previous works done predicting business reviews via decision tree and sentiment identification (and polarity) methods. They present comprehensive guides for data preparation techniques such as stop word removal, and 'Bag of Words', aiding my search for R code in forums which in turn helped to manipulate the corpus text for more accurate results. But, while literature gives insight to which models they apply for predictions, providing source code in conjunction to these methods could further develop my R skills as it could explain the intuitions behind certain features in these commands.

REFERENCES

Shafique, U. and Qaiser, H., 2014. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), pp.217-222.