Week 4 Reflection

Adam Behun

3 Lessons and 1 Business Use case

**Amount of information - Information Entropy**

This week I have learned that the amount of information is a measure of how much reduction in uncertainty a message provides. Amount of information is measured in bits and the more uncertainty it removes, the higher the amount of information. This is very useful not just for the algorithm building decision trees but also a great concept to understand and apply to analysis in general. (<https://www.khanacademy.org/computing/computer-science/informationtheory/moderninfotheory/v/information-entropy>)

**Decision tree**

Decision trees use Entropy to split the dataset based on which feature influences the dependent variable the most (the information that sends the highest amount information = reduces the uncertainty the most). The algorithm continues recursively for each child node until purity - all the instances in that node belong to the same class (categorical values) or have the same target value (numerical values). The goal of the algorithm is to maximize the reduction of entropy (uncertainty) of the dependent variable after each split.

**Problem of collinearity**

This means that one of the independent variables have too strong of an impact on the dependent variable. This makes our model less robust and our predictions less accurate.

**Ensemble Methods**

Ensemble methods are ML algorithms that help us more accurately predict the future by combining the predictions of multiple base models. Thus far, I have learned about Random forests which are an ensemble of multiple decision trees.

**Random forest**

Random forests help us solve the problem of collinearity because using them, we do not have one tree but a collection of trees with a large number of different predictions for each row. These predictions get averaged in numerical problems or voted for in classification problems (how many predictions predict this row to be of class a, b, or c? 🡪 majority wins). This way we solve the problem of collinearity.

**How will this week’s topic help you make better business decisions?**

Knowledge of decision trees and random forests is proved to be useful, yet I consider learning about Information Entropy and the theory behind the amount of information very useful. I think it is applicable to all areas of life, especially for me as a person trying to make useful observations from a high volume and complex datasets. Even partial knowledge of what information decreases the uncertainty in a dataset the most can have a big impact on my observations.