Week 2 Reflection

Adam Behun

**ML with Linear regression**

Linear regressions on a graph are straight lines which makes it easy to visualize how we can predict its future trajectory and points. Equation for LR: y = a + bx. In regression problems we can only use numerical variables (if we have any categorical values, we have to convert them to numerical - preprocessing). Independent variables are the ones that have influence on our dependent variable - dependent variable is the one we are trying to predict. If we have a dependent variable that we are trying to predict, we need to have a training dataset which has values for the dependent variable as well and then we fit an ML model using the method of least squares onto that training dataset. After that, the model finds the relationship between the dependent and independent variables and creates a new column with the predicted values (Predicted using LR). It calculates the R-Squared which shows us how well the model fits our data. We could calculate all this manually, but it is not scalable or efficient 🡪 in that way ML is an automation and helps us speed up the process of predicting future values with levels of certainty.

**Interpretation of results from R and statsbuddy**

Once we prepare our dataset and decide to use LR model, we can use statsbuddy to generate the R code and R-studio to do computations for us. These are the results we will receive (not an exhaustive list)

- .csv file with predicted using LR column - this column shows us the predicted values for the dependent variable using LR

- R-squared value - shows how much of the variance in dependent variable is explained by the independent variables

- Significance codes - show us which independent variables impact the dependent variable the most

**Evaluating the LR model's accuracy**

The Method of Least Squares use with linear regression problems is a technique used to minimize the sum of vertical offsets making sure that our data fits the linear regression line as closely as possible - it calculates the optimal coefficients for the equation. In other words, we are trying to make sure that there is a minimum distance between our data points and the model's line. Another way to evaluate our prediction is to use the R-squared value. R-squared in linear regression models is a measure used to check how well the model fits the data. It ranges from 0 to 1. A 12% R-squared score means that our model explains 12% of the variability in our dependent variable.

**Extra notes:**

Another advantage of using ML to find patterns in data and not visualizations is that visualizations cannot take as many variables as ML models can. We can let the computer understand the relationships and the significance of those relationships between more than 3 variables.

**Business Use case:**

There are many business use cases for machine learning models. One of these use cases is credit risk analysis where financial institutions are trying to identify customers that are going to default on a loan. They can use the past data about customers that defaulted on a loan. Income, credit score, age, and other variables as their independent variables and use the amount owed as the dependent variable. Once they identify a pattern and fit a model using the method of least squares, they can see the predicted values of amount owed for the future, potential customers. They can then evaluate the results to decide how likely a particular prospect is to default on a loan.