Course Project

Professor Corey

Junyu Li, Lanjun Wang

Mar 3rd, 2025

Phase One

A screenshot of a computer screen

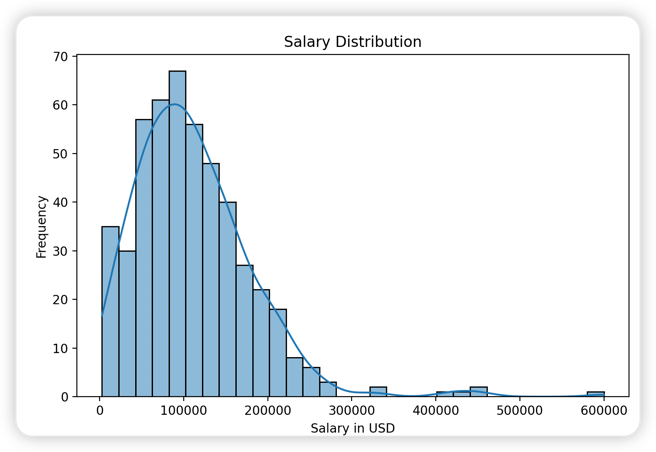
Description automatically generatedFor the binary classification dataset, I chose the Breast Cancer Wisconsin (Diagnostic) Dataset. The dataset is from the Kaggle website as a public dataset. The dataset contains 455 samples and 32 features. All features are numeric. The target variable (diagnosis) is Malignant (1) and Benign (0).

Before conducting any analysis, I divided the dataset into two parts: a training set and a testing set. The ratio of training and testing sets is split to 0.8 to 0.2. The ratio of the target variable between 1 and 0 in both sets is similar. In the training set, the ratio between 0 and 1 is 0.626 vs 0.374. For the testing set, the ratio is 0.631 vs 0.369. After splitting, the feature scaling was performed on the training set to prevent any bias. There is no missing variable in the dataset.

A screenshot of a computer screen

Description automatically generatedThe features contain all the size details of the cancer area. After doing some exploratory data analysis, I found out that the dataset has a clear distribution between the malignant and benign. There is no clear hyperplane that can separate the group into two, due to the middle overlap and sometimes there is some noise in the other group of the set. It will make the training harder. Also, I found out that in some features, two groups almost entirely overlap with each other, making it harder to train on.

By seeing the heapmap, we can find some features have a strong relationship to different features, but some are completely unrelated. It also makes the training more difficult due to the fact that not all data is related to each other.

For the classification dataset, I chose the Data Science Job Salaries Dataset. The dataset is from the Kaggle website as a public dataset. The dataset contains 606 samples and 12 features. In 12 features, numerical features have 5, ordinal features have 1, binary features have 1, and categorical features have 5. The target value is set to salary\_in\_usd to predict the salary in data science. Before the splitting, first I deleted the index column for the samples, then I assigned integers to the ordinal column, and lastly I performed the one-hot encoding for the categorical features. After spliting, the ratio between training set and testing set is 0.79:0.21. There is no missing variable across the datatset.

For the training set, it has 485 samples. The mean is 111050. Standard deviation is 70592. Min is 2859. Max is 600000. Based on the salary distribution, the majority of samples’ salaries are around the 0 – 200000. The entry-level sample is 14.50%, mid-level sample is 35.09%, senior-level sample is 46.13%, and executive is 4.28%.

A graph with red lines and blue squares

Description automatically generatedBy looking at the heatmap, it is clearly understood that there is so little relation between the features. The dark red (strong relationship) in the picture was the relation between the currency company using and the company location.

There are some challenges in the dataset: First, as we can see in the heatmap, there is no relationship between features. It causes issues to machine learning that it might not have enough common to learn from it and predict the trend. Second, the features might be too little that it doesn’t contain lots of useful features like the binary dataset. Third, some features need more processing. There is a feature called “job\_title”. It will create lots of issues to group the information in the features. There is a chance that two different job titles are doing the exact same job. It will make the machine learning harder because it is hard to determine if two different jobs mean the same thing. It can make the result inaccurate if there is miss grouping for different job titles.