Course Project Phase 1
Robbie & Parsa

Binary Dataset: ISOT Fake News Detection Dataset

The Fake News Detection Dataset is a binary dataset that attempts to predict whether a given news article is “real” news or “fake” news. It uses contents of the article and patterns related to headlines or statements to decide the validity of the information given in the article, then labels it as either “real” or “fake”. The dataset contains a total of 44,919 instances, with 21,417 “real” articles and 23,502 “fake” articles, and four features that consist of information related to the article. The title is the headline of the article, which may either be misleading or outright false, likely to be the biggest indicator of whether an article is real or fake news. The text contains the body of the article and expands on the headline given in the title, but may be less false or less misleading than the title. There are also the subject of the article (e.g. “news”, “politics”, etc.) and the date it was published (ranging from March 30, 2015 and February 18, 2018), however these are likely to have minimal impact on whether an article is real or fake news. There is minimal missing data, with around 3% instances not having an article body, and there are also some mismatched dates in the fake dataset, although this will likely not have a noticeable impact.

Due to the large size of the dataset, we expect that it may be time consuming to train our models, especially if we test multiple types of models with different hyperparemeters. One way to work around this would be to train each type of model on a subset of the training data we select, then decide which model we want to use to train the entire dataset on. On the other hand, having a large dataset helps to minimize the impact of outliers or noise that might show up in the dataset and allow us to develop a model that gives an accurate representation of real versus fake news articles. In addition, our labels are roughly evenly split, so we should not run into issues related to misleading accuracy or precision scores.

The dataset is split into two CSV files, one containing “real” news articles and the other containing “fake” news articles. We will take each of these files and split them into 80% training and 20% testing, then combine each training and testing dataset with the appropriate labels so that we have one training set and one testing set. We chose an 80/20 split because our dataset is large, so we will be able to train on a larger amount of data than a 70/30 split while still having sufficient data remaining for testing.

1. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Multiclass Classification Dataset: The Customer Segmentation Dataset

The Customer Segmentation dataset is provided by an automobile company as part of an effort to optimize outreach by segmenting their customers. The primary task with this dataset is to classify each individual into one of four possible customer segments (A, B, C, or D) using a variety of personal and behavioral features. There are a total of 2,627 examples, each representing a potential or active customer.

The dataset's features include customer ID, gender, marital status, age, graduation status, profession, work experience in years, spending score, family size, and a categorical variable labeled Var_1. The gender feature is split approximately evenly between male (54%) and female (46%). Marital status is recorded as true for 58% of instances, false for 40%, with about 2% missing values. The age range spans from 18 to 89, with most individuals falling into the 25-55 year-old category. Graduation status is marked true for around 61% of examples, false for 38%, and 1% are missing. Profession is a categorical feature, with Artist and Healthcare being notable values; Artist comprises about 31% and Healthcare about 16% of the dataset, while all the other professions make up the remainder. Work experience ranges from zero up to fourteen years, but the majority of customers report low levels of experience, typically around zero or one year. The spending score is divided into categories, with the lowest category making up approximately 62% of entries, the average group accounting for 24%, and all other categories representing 14%. Family size ranges from very small (one) to fairly large (nine), but most families are small to moderate in size. The variable Var_1 is categorical as well, with Cat_6 being the most common (64%), Cat_4 accounting for 15%, and all other categories are up to 21%.

A few challenges with the dataset are evident. Some features, such as marital status and graduation status, contain missing or null values. Categorical features like profession and Var_1 are imbalanced, meaning one category makes up the majority of cases. Appropriate preprocessing will be required to address missing data.

The dataset comes split into train and test sets. The training set is intended for building and evaluating models, while the test set is reserved for some final evaluations. For data analysis, key insights may include basic descriptive statistics such as mean and median for age, work experience, and family size. Missing values should be specifically analyzed to quantify the extent of missingness for each feature.

Finally, an examination of customer segment labels will allow verification of how balanced the multiclass distribution is across A, B, C, and D.

1. https://www.kaggle.com/datasets/abisheksudarshan/customer-segmentation