**Outline of Group Presentation**

**Choosing a Dataset**

We were given three datasets to review for our group project. The first step was to take review each dataset and the corresponding data dictionaries/README files to understand what we had.

**Dataset 1 – Bank Marketing**

The dataset consists of 17 variables, of which 7 are numerical / quantitative. Of the available datasets, this one had the most total data—more than 45,000 rows and 17 features, with almost no missing data points. It also has a good mix of quantitative and categorical variables, and the data dictionary was very useful describing both the features and the overall purpose of the dataset.

**Dataset 2 – Housing Data**

The dataset includes data for houses sold over a 1-year period, from May 2014 to May 2015, and consists of more than 21,000 rows with 21 features. The most obvious open question was to ask, can we predict housing prices based on the given data features? This would most likely be a multivariate regression model; however, we weren’t very interested in this question.

**Dataset 3 – Online Shoppers**

This consists of more than 12,000 rows with 18 features. We rejected this dataset almost immediately for a number of reasons. In no particular order:

* The corresponding data dictionary was rather unhelpful in describing both the details of feature variables, as well as the overall purpose of the full dataset
* The distribution of categorical variables didn’t make much sense—for example, more than 5000 records between March and May but 0 records in the month of April
* Numerical variables tended to have long-skew distributions, and there were many imbalanced data classes for categorical variables plus lots of zeroes

Thus, we chose Dataset #1 because it had the most data to work with, a good mix of categorical and quantitative variables with relatively few missing data points, and a good data dictionary/README description to understand the purpose of this dataset (predicting deposits).

**Exploratory Data Analysis**

The Bank Marketing dataset is based on the marketing campaigns of a Portuguese banking institution, with observations recorded from May 2008 to November 2010. The purpose of these campaigns was to encourage potential clients to sign up for a bank term deposit, which was successfully recorded for 11.7% of all observations.

**Structure of the Dataset**

The first step was to quickly understand the available variables within the dataset. Pulling the CSV file into a Pandas (Python) dataframe and taking a quick look at the first several rows:

A picture containing icon

Description automatically generated

For numerical quantities, we can also take a quick look at the summary statistics—mean, standard deviation, IQR, etc.

Table

Description automatically generated

It should be noted, however, we later converted **day** to a categorical variable, because itrefers to *day of the month* that a client was contacted, according to the data dictionary, and should not be treated as a numerical or ordinal variable.

We also later replaced **pdays** and **previous** with a binary Yes/No variable, as will be described later.

**Age**

We plotted histograms for the **age** variable, and the distribution at least “makes sense” in that there seems to be a continuous probability distribution:

Chart, histogram

Description automatically generated

As another sanity check, we created another histogram and boxplot separated by **marital** status. The findings align with common sense understanding, namely that single people tend to be younger and divorced people tend to be older—after all, you can get divorced before you get married.

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

However, the distribution for **age** does not seem to change much when separated by **deposit** status. That is to say, age does not seem to be a reliable indicator for whether a person is more or less likely to sign up for a bank term deposit.

A picture containing text, music

Description automatically generated

Chart, box and whisker chart

Description automatically generated

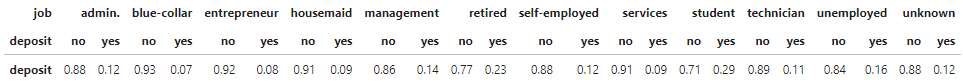
**Job**

When reviewing the **job** variable, there seems to be some variation in proportions of Yes/No for **deposit**, although the sample sizes do get rather small and thus may be less reliable indicators for certain **job** categories.

Chart, bar chart

Description automatically generated

Percentages:



**Marital**

A change in **marital** status doesn’t seem to have a dramatic change in **deposit** status.

Chart

Description automatically generated

Text, table

Description automatically generated

**Education**

A change in **education** level doesn’t seem to have a dramatic effect on **deposit** status.

Chart

Description automatically generated

**Default**

This refers to whether or not a potential client has credit in default, and is marked by a binary Yes/No. While there does seem to be a difference in **deposit** status rate based on **default** status, the overall **default** classes are so imbalanced that this may be a small sample size difference.

Graphical user interface

Description automatically generated with low confidence

A picture containing icon

Description automatically generated

**Balance**

The **balance** variable is both quantitative and continuous, and it represents the potential client’s average yearly balance in Euros. While the full range goes from -8019 to +102127 Euros, for visualization purposes we restrict the plots to a range of -3000 to +20000 Euros.

Shape

Description automatically generated

If we look instead at the boxplot for **balance** separated by **deposit** status, we see that the IQR is shifted to the right for those who opened a bank term deposit. In other words, people who opened a bank term deposit tend to have a higher bank balance than those who do not—and this aligns with a common sense idea that people with more money are more likely to enjoy certain financial instruments.

Chart, box and whisker chart

Description automatically generated

It’s also worth noting that if we instead separate the boxplot by **default** status, we see that the people most likely to default tend to have a negative yearly **balance**. And this also aligns with a common sense notion, that people who owe money are more likely to default on their financial obligations than people who do not owe money.

Chart, box and whisker chart

Description automatically generated

**Housing**

When looking at the binary variable for **housing**, it seems like a true statement to say that people who do not have a **housing** loan are more than twice as likely (on a proportional basis) to agree to a bank term **deposit**.

Chart

Description automatically generated

Table

Description automatically generated

**Loan**

The yes/no status of a personal **loan** seems to have an impact on whether a potential client agrees to a bank term **deposit**, similar to the **housing** variable, although for personal **loan** status the binary yes/no classes are more imbalanced and thus may be more prone to small sample errors.

Chart, bar chart

Description automatically generated

Table

Description automatically generated

**Contact**

The **contact** variable consists of 3 categories, one of which is *unknown*. For those potential clients that we do know the contact method, whether *cellular* or *telephone*, there doesn’t seem to be a significant difference in **deposit** status proportion rates.

Chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

**Day**

The **day** variable refers to day of the month that a potential client was contacted, and was originally recorded as a numerical variable. However, because of the periodic nature for days of the month, it makes sense to recode this as a categorical variable—otherwise the model may give a weight to day 20 which is 10x higher than the weight for day 2, similar to linear regression models. For additional views, we looked at stacked histograms and filled plots to get a visual sense for how **deposit** status may change with day of the month.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**Month**

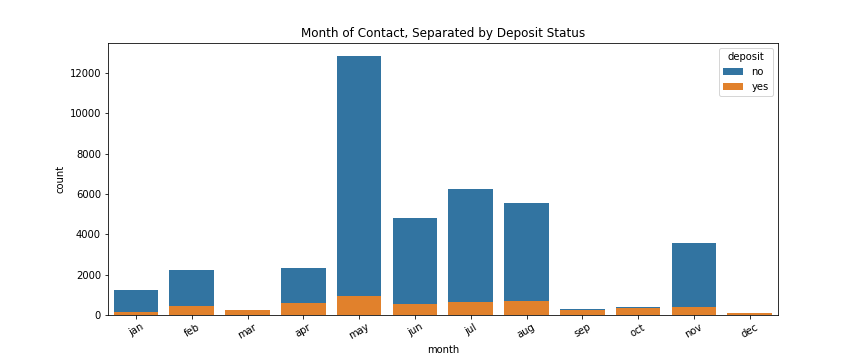
The **deposit** status varies quite a bit my **month**, and indeed the **month** classes themselves are not uniformly distributed. While we do know the study took place from May 2008 to November 2010, and thus the months from May to November should be disproportionately represented, it is unknown why the actual distribution varies so widely between months and whether that affects **deposit** status.

Chart, bar chart

Description automatically generated

**Duration**

The **duration** variable is a measurement, in seconds, of how long the last contact lasted. When we look at the boxplot separated by **deposit** status, it’s clear that the IQR for *yes* is shifted to longer times than the IQR for *no*, which makes sense that people who secure a bank term deposit are more likely to have a longer conversation than people who simply say no. However, for modeling purposes we’ll need to be very careful with this variable—if we want to predict which customers are more likely to sign up, such that we can save resources by contacting fewer people, we may not actually know the **duration** variable before contacting potential clients.

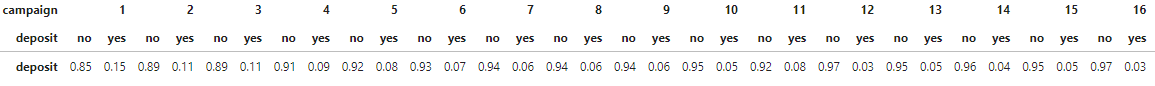
**

**Campaign**

The **contact** variable measures how many contacts were performed during a campaign for a given client. According to the data, those who agreed to a bank term **deposit** where more likely to say *yes* in the first few contacts rather than multiple contacts. Perhaps we should consider a cutoff of 15-20 contacts per person to eliminate excessive calls which drain time and resources for little return.

Shape, rectangle

Description automatically generated



**pdays** and **previous**

The **pdays** variable refers to the number of days that passed since the client was last contacted by a previous campaign (no contact is recorded as -1), and the **previous** variable refers to the number of contact performed before this campaign and for this client (no contact is recorded as 0). However, the count of (-1) for **pdays** is exactly the same as the count of 0 for **previous**, which is 39922 observations. Because this represents more than 75% of the recorded observations, it makes sense to combine these into a single binary variable which flags whether or not the potential client was *ever* contacted previously for a different campaign.

Chart, bar chart

Description automatically generated

Initial Regression Model

Including Multiple Models for Comparison

Imputing Data

Different Scalers