Three for two

Jonathan Ferdinand, Devina Gera, Archimedes Li, Henry Liu,
Katherine Shi, Sam Stevens

April 1, 2025

1 Introduction

Our model is a generalized linear regression model that uses the logit link function to perform a binary classification of raisins to either the Kecimen or Besni class. Our model uses seven independent variables, further detailed in the data description, to predict a probability as following:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{7} \beta_i X_i$$

where we solve for p, the probability that a raisin belongs to the Besni class. Our model then compares the probabilities p and $1-p$, classifying the instance as Kecimen if $1-p > p$ and Besni otherwise.

We choose the logit function as our link function due to its conversion of a problem involving linear parameters to a binary classification output by modeling probabilities for two classes. Our model, as with most linear models, lacks complexity as opposed to an alternative solution, such as one using neural networks. As such, we chose a dataset with simple parameters that have a linear relation to the output to amend for this fact.

The R packages that we used were `caTools, tidyverse, caret, ggplot2, dplyr, pROC, ggfortify`

Our model was trained on [XXX] instances and performed with [ACCURACY]

2 Data Description

Our dataset is for raisin classification between Kecimen and Besni raisin types. The dataset extracts 7 different features from 900 images of raisins, with 450 of each type of raisin. Area - Gives the number of pixels within the boundaries of the raisin. Major axis length - Gives the pixel length of the main axis, which is the longest line that can be drawn on the raisin. Minor axis length - Gives the pixel length of the small axis, which is the shortest line that can be drawn on the raisin. Eccentricity - It gives a measure of the eccentricity of the

ellipse, which has the same moments as raisins. Values closer to 0 indicate the raisin is more circular, and values closer to 1 indicate that the raisin is more elongated. Convex area - Gives the number of pixels of the smallest convex shell of the region formed by the raisin. Extent - Gives the ratio of the region formed by the raisin to the total pixels in the bounding box. Ranges from 0 to 1. Perimeter - It measures the environment by calculating the distance between the boundaries of the raisin and the pixels around it.

3 Analysis

The data contains two-hundred samples of advertised products and three features (`TV`, `radio`, and `newspaper`) and one variable (`sales`) for each product. Variables `TV`, `radio`, and `newspaper` indicate the advertising budget for TV, radio, and newspaper, respectively. `sales` indicate the number of sales of the product[1].

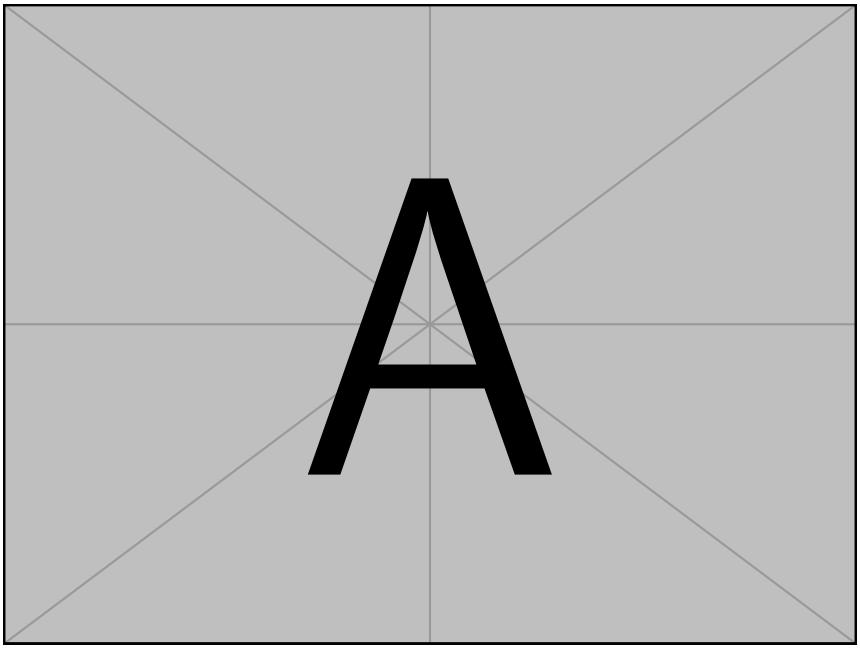

Figure 1: `TV` and `radio` areslightly skew right, while (`sales` is normally distributed. `newspaper`) is extremely skew right.

The `newspaper` column clearly has an outlier in the histogram, which we remove. The data passed our tests for linearity (See Analysis Section), so no additional pruning was necessary.

[1] https://search.r-project.org/CRAN/refmans/glmtoolbox/html/advertising.html

4 References

We used the `Advertising.csv` dataset from Trevor Hastie's "An Introduction to Statistical Learning" GitHub page[2].

[2] https://trevorhastie.github.io/ISLR/Advertising.csv