Final_project

Connor Demorest

11/28/2020

Data Analysis (Scaled to be worth 26 points)

Using the Indian recipe dataset fit a logistic regression model to model the probability of a dish being classified as a main course. Write your results in a short report (shorter than the projects). Turn this document in separately. Including figures and tables, I am setting a four page maximum using standard PDF output settings in RMD. This will require careful selection and sizing of tables and figures. The page limit does not apply to references or code in the appendix.

Report generalities	**Points**
Spelling, grammar, writing clarity, paragraphs, section labels	/8
Citations/Acknowledgments for papers and packages used	/4
Code in appendix	/4

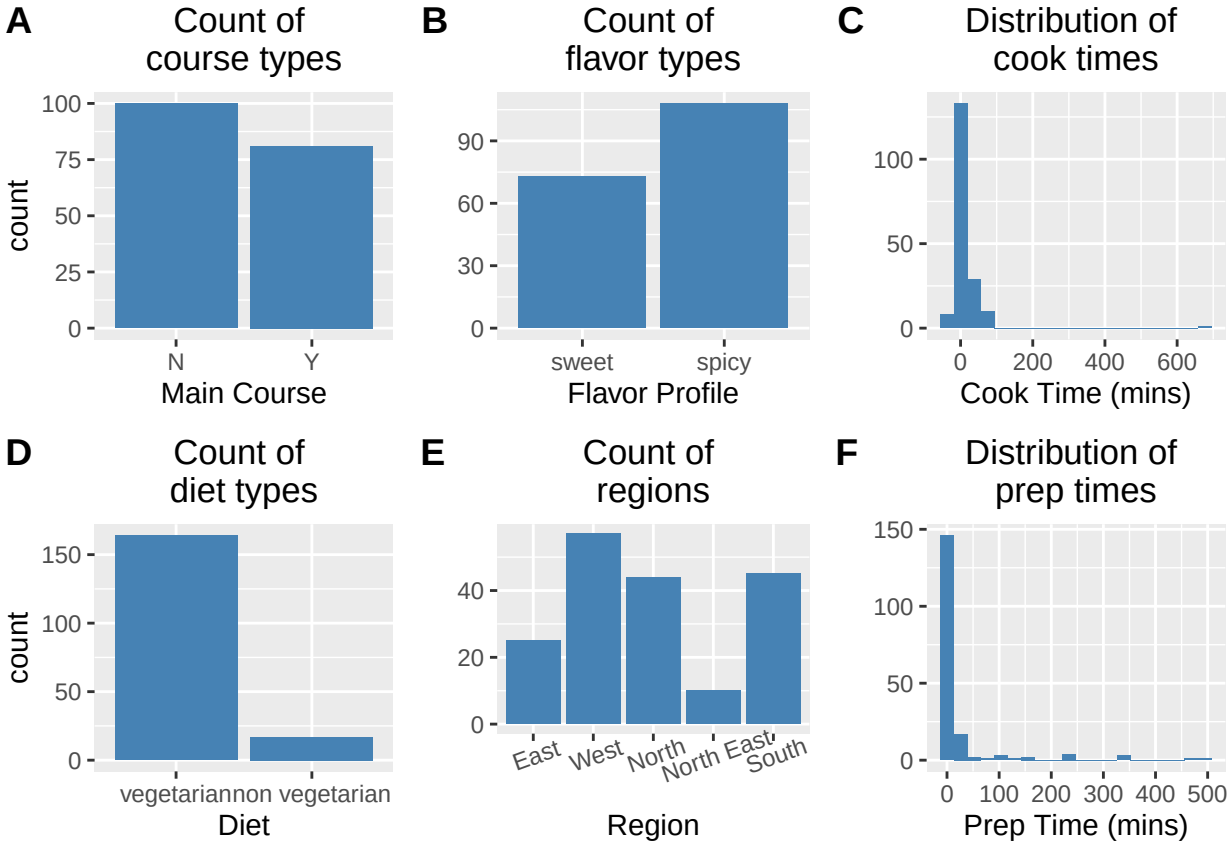
Introduction + Data Overview	**Points**
Research question	/4
Variables with units and descriptive statistics	/4
Data Viz: Figure Clarity (Titles, Labels, and Captions)	/4

Statistical Procedures	**Points**
Define model to fit with complete notation (including priors)	/8
Defense of model choice	/4

Results + Discussion	**Points**
Discuss Results in the context of the research question	/4
Summarize estimates from final model including uncertainty	/8

Introduction & Data Overview

The data contains information about characteristics of Indian foods. Figures A through F shows basic summary statistics of the outcome variable and some of the predictors. The research question is to predict if a food is a main course based on characteristics available in the dataset.



The data are available on Github. From that original data set, there were 3 changes: a new variable called "Main Course" was created that indicates "Y" if the food is a main course and "N" if the food is not a main course and the prep and cook times were centered to the median prep and cook times, and the ingredient variable was removed because it was not used.

Statistical Procedures

The model being fit is a logistic regression using the following predictors: diet (vegetarian orot vegetarian), region (N, S, W, E, and NE), prep time (mins), cook time (mins), and flavor profile (spicy or sweet) to predict the outcome variable of main course (Y or N). There are 3 candidate models considered.

The first model is a basic model with no interaction terms:

$\widehat{Main\ Course}(Y\ or\ N) = diet + prep\ time + cook\ time + flavor\ profile + region$

Interactions between prep time and flavor and cook time and flavor were then examined and both are statistically significant (p-value: prep time * flavor = 0.002, p-value: cook time * flavor = 0.024), so model 2 with those interactions was considered.

$\widehat{MainCourse} = prep\ time + cook\ time + flavor\ profile + diet + region + prep\ time * flavor = spicy + cook\ time * flavor = spicy$

The third model removes diet as a predictor because it was not statistically significant as a predictor in model 2 (p = 0.660).

$\widehat{MainCourse} = prep\ time + cook\ time + flavor\ profile + region + prep\ time : flavor = spicy + cook\ time : flavor = spicy$

The best model was chosen by the lowest AIC of the three candidate models. The model that performed the best was mnodel 3, as shown in Table 1.

Table 1: AIC of three candidate models

Model	AIC
Model 1	137.6
Model 2	131.1
Model 3	129.3

The estimates for the coefficients of the predictors in the best model is shown in Table 2, with the estimated β values for the the final fitted model:

$logOdds(Main\ Course = Yes) = \beta_0 + \beta_1 * PrepTime + \beta_2 * CookTime + \beta_3 * Flavor = Spicy + \beta_4 * Region = West + \beta_5 * Region = North + \beta_6 * Region = NorthEast + \beta_7 * Region = South + \beta_8 * PrepTime : Flavor = Spicy + \beta_9 * CookTime : Flavor = Spicy + \epsilon$

Or, alternatively,

$Odds(Main\ Course = Yes) = exp(\beta_0 + \beta_1 * PrepTime + \beta_2 * CookTime + \beta_3 * Flavor = Spicy + \beta_4 * Region = West + \beta_5 * Region = North + \beta_6 * Region = NorthEast + \beta_7 * Region = South + \beta_8 * PrepTime : Flavor = Spicy + \beta_9 * CookTime : Flavor = Spicy + \epsilon)$

Table 2: Estimated coefficients of the predictors for model 3

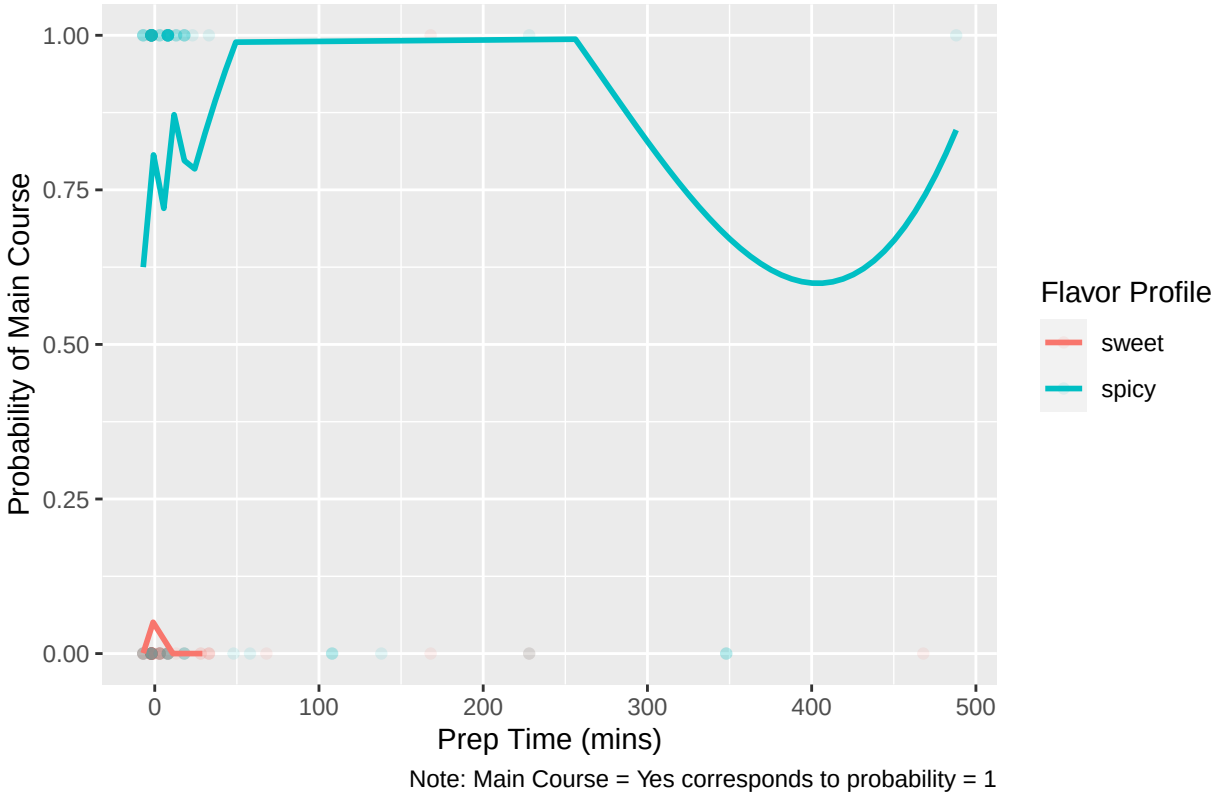
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.925	0.950	-3.079	0.002
`Prep Time`	0.003	0.005	0.662	0.508
`Cook Time`	0.003	0.009	0.298	0.766
`Flavor Profile`spicy	5.651	0.956	5.912	0.000
regionWest	-2.560	1.181	-2.168	0.030
regionNorth	0.707	1.187	0.596	0.551
regionNorth East	0.081	1.547	0.052	0.958
regionSouth	-1.177	1.182	-0.996	0.319
`Prep Time:Flavor Profile`spicy	-0.022	0.007	-3.085	0.002
`Cook Time:Flavor Profile`spicy	0.050	0.022	2.257	0.024

The probability that a dish is a main course is:

$P(Main\ Course = Yes) = \frac{Odds(Main\ Course=Yes)}{1+Odds(Main\ Course=Yes)}$

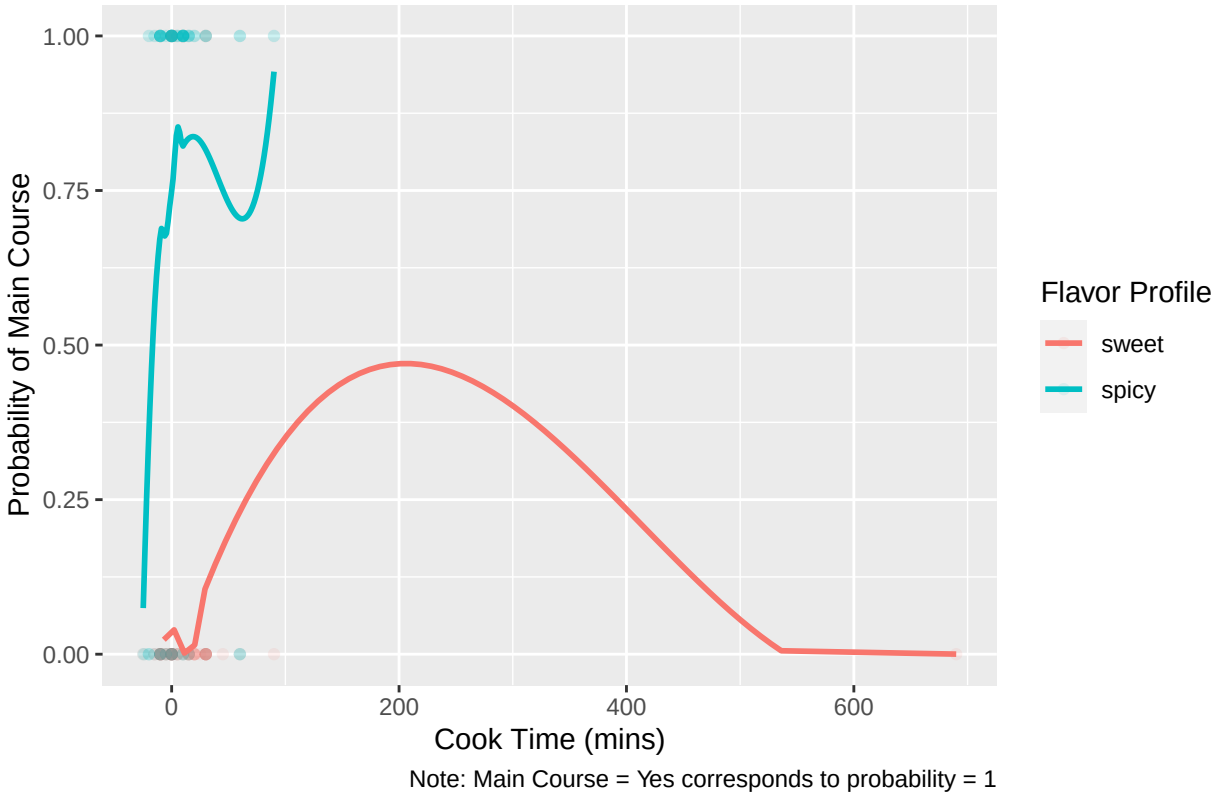
and is shown in Figures G and H.

Figure G: Probablity of main course by prep time and flavor



Note: Main Course = Yes corresponds to probability = 1

Figure H: Probablity of main course by cook time and flavor



Note: Main Course = Yes corresponds to probability = 1

Figures G and H show the probability of a dish being a main course by the cook time and the flavor profile. They look weird, I think because there are not many main course dishes that are sweet. If there were more data I think this might sort itself out and look more like a logistic curve possibly. As it is I don't know if you really gain any information from these figures aside from that the curves look very different between the flavor profile, which indicates that there may be an interaction effect going on.

Results & Discussion

The most important predictor seems to be if the food is spicy or sweet. If the food is spicy, the log-odds of a food being spicy increase by 5.65 times, or a around 284 times higher odds of a food being a main course if it is spicy ($p < 0.001$, 95% CI: 42.1, 1925.6), when holding everything else constant.

Another interesting thing is that the prep time and cook time are only meaningfully important for the interaction between a food being spicy and a the time taken to cook or prep the food. If the food is sweet, the estimated effect of a 1 minute change in cook or prep time is associated with less than a 1% change in odds of a food being a main course when holding all other predictors the same. If a food is spicy, a 1 minute increase in prep time is associated with a 2.2% decrease in odds of the food being a main course ($p = 0.002$, 95% CI: 3.5, 0.8) and a 1 minute increase in cook time is associated with a 5.1% increase in odds of a food being a main course ($p = 0.024$, 95% CI: 0.6, 9.9).

References

https://stackoverflow.com/questions/31182147/how-to-suppress-automatic-table-name-and-number-in-an-rmd-file-using-xtable-or for how to fix the automatic captioning of the tables in knitr and Latex.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Dominic Comtois (2020). summarytools: Tools to Quickly and Neatly Summarize Data. R package version 0.9.6. https://CRAN.R-project.org/package=summarytools

Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.0. https://CRAN.R-project.org/package=cowplot

Appendix

```
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center', message = F, warning = F)

library(tidyverse)
library(summarytools)
library(cowplot)
library(knitr)


food = read_csv('https://raw.githubusercontent.com/stat408/final_exam/master/indian.csv', col_types = "
```

```
  mutate("Main Course" = factor(ifelse(course == "main course", "Y","N")),
         "Prep Time" = prep_time - median(prep_time),
         "Cook Time" = cook_time - median(cook_time),
         "Flavor Profile" = flavor_profile) %>%
  select(-prep_time, -cook_time, -flavor_profile, -ingredients, -course)

# Plots of counts for each variable

f1 = ggplot(data = food, mapping = aes(x = `Main Course`)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Count of \ncourse types") +
  theme(plot.title = element_text(hjust = 0.5))
f2 = ggplot(data = food, mapping = aes(x = `Prep Time`)) +
  geom_histogram(fill = "steelblue", bins = 20) +
  labs(x = "Prep Time (mins)",
       y = "",
       title = "Distribution of \nprep times") +
  theme(plot.title = element_text(hjust = 0.5))
f3 = ggplot(data = food, mapping = aes(x = `Cook Time`)) +
  geom_histogram(fill = "steelblue", bins = 20) +
  labs(x = "Cook Time (mins)",
       y = "",
       title = "Distribution of \ncook times") +
  theme(plot.title = element_text(hjust = 0.5))
f4 = ggplot(data = food, mapping = aes(x = diet)) +
  geom_bar(fill = "steelblue") +
  labs(x = "Diet",
       title = "Count of \ndiet types") +
  theme(plot.title = element_text(hjust = 0.5))
f5 = ggplot(data = food, mapping = aes(x = region)) +
  geom_bar(fill = "steelblue") +
  labs(x = "Region",
       y = "",
       title = "Count of \nregions") +
  theme(axis.text.x = element_text(angle = 20, hjust = 0.5),
        plot.title = element_text(hjust = 0.5))
f6 = ggplot(data = food, mapping = aes(x = `Flavor Profile`)) +
  geom_bar(fill = "steelblue") +
  labs(y = "",
       title = "Count of \nflavor types") +
  theme(plot.title = element_text(hjust = 0.5))
cowplot::plot_grid(f1,f6,f3,f4,f5,f2, labels = LETTERS[1:6])


model1 = glm(data = food, family = "binomial", formula = `Main Course` ~ diet + `Prep Time` + `Cook Time
aic1 = model1$aic

model2 = glm(data = food, family = "binomial", formula = `Main Course` ~ (`Prep Time` + `Cook Time`) *
aic2 = model2$aic

model3 = glm(data = food, family = "binomial", formula = `Main Course` ~ (`Prep Time` + `Cook Time`) *
aic3 = model3$aic
```

```r
aic_df = data.frame(Model = c("Model 1", "Model 2", "Model 3"), AIC = c(aic1, aic2, aic3))
kable(aic_df, digits = 1, caption = "Table 1: AIC of three candidate models")

kable(summary(model3)$coefficients, digits = 3, caption = "Table 2: Estimated coefficients of the predi


food2 = food %>% mutate(main_course = ifelse(`Main Course` == "Y", 1, 0))

ggplot(data = food2,
       mapping = aes(x = `Prep Time`,
                     y = main_course,
                     col = `Flavor Profile`)) +
  geom_point(alpha = 0.1) +
  geom_smooth(formula = "y~x", method = "loess", se = F) +
  labs(y = "Probability of Main Course",
       title = "Figure G: Probablity of main course by prep time and flavor",
       x = "Prep Time (mins)",
       caption = "Note: Main Course = Yes corresponds to probability = 1") +
  ylim(0, 1) +
  theme(plot.title = element_text(hjust = 0.5))

ggplot(data = food2,
       mapping = aes(x = `Cook Time`,
                     y = main_course,
                     col = `Flavor Profile`)) +
  geom_point(alpha = 0.1) +
  geom_smooth(formula = "y~x", method = "loess", se = F) +
  labs(y = "Probability of Main Course",
       title = "Figure H: Probablity of main course by cook time and flavor",
       x = "Cook Time (mins)",
       caption = "Note: Main Course = Yes corresponds to probability = 1") +
  ylim(0,1) +
  theme(plot.title = element_text(hjust = 0.5))
```