Homework 12 Report

Chenning Xu

1 Summary Staistics

Table 1: Summary statistics - continous variables

	count	mean	sd	min	max
Top 20% Chef Indicator	3000	.2633333	.4405151	0	1
Age (Years)	3000	35.01733	9.544321	18	65
Years of Experience	3000	10.03767	4.886465	0	31
Knife Skills (1-10)	3000	5.927597	1.981052	1	10
Plating Aesthetics (1-10)	3000	6.990127	1.439236	1.04	10
Creativity (1-10)	3000	7.86127	1.718801	1.71	10
Challenge Win Rate (%)	3000	19.98385	9.773741	0	52.14
Judges' Feedback (1-10)	3000	6.496	.9686861	3.32	9.87
Stress Management (1-10)	3000	5.07865	2.625114	1	10
Social Media Following (0-10)	3000	5.03166	1.98971	0	10
Audience Popularity (0-10)	3000	4.97524	2.881288	0	10
Signature Dishes Created	3000	3.003667	1.764849	0	10
Unique Ingredients Used	3000	5.039333	2.203952	0	14
Weekly Practice Hours	3000	20.16535	5.059817	0	37.26

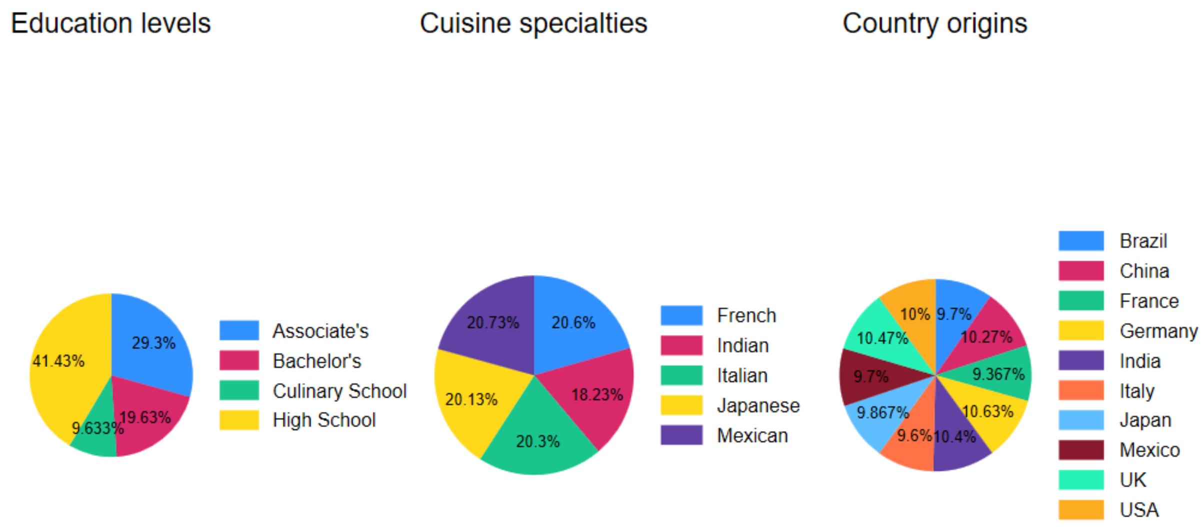

Figure 1: Categorical characteristics of Chefs

2 Models and Performamce

I tested three models:

Model 1: A logistic regression model using all continuous and categorical variables to predict the probability of a chef being in the top 20%. This serves as the benchmark model.

Model 2: A lasso regression model, which incorporates all variables and their interactions as the initial input.

Model 3: A simplified model with manually selected variables. In this model, I excluded variables such as the number of signature dishes, weekly hours practiced, and all the categorical variables. These were omitted because I found no apparent relationship between them and the target variable based on LPOLY graphs. Additionally, this model does not include interactions or log transformations of the independent variables, as manual inspections did not reveal significant relationships warranting such transformations. However, this approach carries the risk of overlooking interactions that, while not intuitively obvious, might possess predictive power.

The evaluation matrix of the three models is presented below.

	Model 1	Model 2	Model 3	Best
Log-Likelihood	-1091	-1086	-1098	Model 2
Pseudo-R2	0.3693	0.3721	0.3649	Model 2
AUC	0.8832	0.8863	0.8808	Model 2
Classification Accuracy(50%)	0.83	0.8323	0.829	Model 2
Number of Variables	30	28	12	Model 3

By any accuracy measure, the second model, which uses lasso, is the best performer. However, its advantage over the other two models appears to be marginal. Given that Model 3 uses less than half the variables of the other two models yet achieves comparable predictive power, I consider Model 3 to be the best approach.