Bella & Bona Data Analysis Case Study

Challenge 2.1

This part includes the calculation process of the KPI “Capture Rate”. The steps taken in calculation are as follows:

- Extracting the number of days from the column “Contractual Weekdays” in the sheet “company data” and adding it as a new column.
- Applying a user + week based groupby to the “user data” sheet to get the number of weekly deliveries to each customer.
- Merging the two modified tables on the “Company Name” column.
- Dividing the number of deliveries to number of contractual days to obtain capture rate.

Then, the groupby step is modified to get monthly, or company-based numbers.

Challenge 2.2

In this part, the existence of a linear relationship between end-user price and capture rate is checked. The correlation coefficient (PCC) is -0.185 for weekly capture rates, and -0.174 for monthly capture rates. This indicates that there is a correlation between end-user price and capture rate, yet it is not that strong (<0.3). As can be expected, the values are negative since end-user price and capture rate are inversely proportional. This is investigated further via visualizations of weekly and monthly rates. The visualizations yielded that the relation is not linear, but offering free meals have a remarkable effect: An additional “Is Free?” column has a coefficient of 0.321. In other words, the end-user price is affects capture rate significantly, *only when the discount is 100%*.

Challenge 2.3

To start with assessing other factors, company sizes and monthly user feedback averages for each rating category are calculated. According to the correlation matrix in the source code, user feedbacks seem like irrelevant, all of which have a magnitude near zero. This probably stems from the fact that many users do not take their times to give ratings, and consequently one-month data becomes inadequate. A possibly useful modification can be narrowing down the scale from 1-20 to 1-5, which would reflect the users' satisfaction better in small sample sizes. On the other hand, company size's coefficient has a higher magnitude compared to the feedbacks, yet its correlation (-0.127) is still weak.

The feature with the highest correlation with capture rate is the number of contractual days, which has a negative coefficient equal to almost 0.6. It is inversely proportional with the capture rate, which can be interpreted as follows: End-users who have fewer opportunities per week to eat B&B meals have a higher tendency to benefit from these opportunities.

At the first glance, offering the most frequent deliveries possible to the businesses might seem like a good idea to maximize profits. Considering the inverse proportionality of capture rate and frequency, reassessing the negotiation strategy might be helpful. For example, an optimum number of weekly deliveries can be determined (such as 2 or 3 days a week), and this number can be used as a first offer.

This would decrease the total profit in the short-term for sure. But as a long-term strategy it can be helpful when supported with a second strategy: Working with more companies. Focusing on increasing the number of businesses served instead of number of days served would probably yield better results. To exemplify*, serving 5 companies 3 days a week would probably be better than serving 3 companies for 5 days a week*, in terms of capture rate, of course.

Challenge 2.4

In this study, two models are built and compared to predict capture rates. Linear Regression yielded a MAPE (Mean Absolute Percentage Error) of 21%, while Random Forest Regressor yielded 8%. The huge performance gap between these two models is reasonable due to two facts: the existence of nonlinear relationships in the data and the size of the data set. As an ensemble method, Random Forest is better in the sense that it takes the most out of a small data set by repeatedly picking random samples from it. The model works best with the following features: “Employee Paid Main Dish Price”, “Num. of Cont. Days”, “Is Free?”, “Rating”, “Satisfaction”, “Taste”.

Challenge 2.5

- "Is Free?" column contributes well. More custom features can be engineered.
- Holidays in Germany and other special days can be introduced as a column. (For example, Ramadan, during which Muslims fast.)
- In addition, separate analysis for each diet type can be held.

Challenge 2.6

Predictions can be found in the source code (Challenge 2, Part VI).

Challenge 3

In this part, a basic exploratory time series analysis is held to get some insights from the data. Due to the small size of the data (only one month), a huge part of its seasonal behavior is obviously lost, such as the effect of seasons, months, and quarters. The only useful time unit in this case is the day of the week. Therefore, rather than a traditional time series forecasting approach like ARIMA or SARIMAX, an ensemble method potentially can perform better. In this study, XGBoost (Extreme Gradient Boosting) is used based on the “day of the week” variable, the forecasting results are as follows:

-	Monday:	394
-	Tuesday:	491
-	Wednesday:	505
-	Thursday:	446
-	Friday:	232