Sold salt- and sugar content comparison to the business goals

Capstone Project Summary for CEU MS in Business Analytics 2022

Gyongyver Kamenar

1. Business problem and project goals

My capstone project work was done for an international retailer company producing and selling several types of products. One of the company's business goal is to decrease the amount of added sugar and salt in their food products because of corporate social responsibility reasons. It is well known that too much added sugar and salt is not good for health so customers should avoid it. The retailer company aims to help health-conscious customers and improve the health of the society with their products, so it plans to decrease the sold quantity of added salt and sugar in their products by 20% until 2025. By reaching this goal, the company would be considered as socially responsible, it would increase its popularity and eventually attract more customers.

To reach their goal the responsible team needs some data science tools to make decisions regarding their products. Based on the results, they might change the recipe of certain products to decrease the amount of added salt and sugar, or they might deduct some products from the shop. To be able to make decisions, they need 3 different metrics.

Firstly, they need a monitoring tool which shows the sold quantity of added salt and sugar since the beginning of the year, on monthly basis. The company has reports on yearly basis, but it is not enough to make regular changes on products. If they make changes during the year, they get the report after the year ends but by then it is too late to act, the year is over. If they would regularly see how they are progressing, they could make better decisions regarding their products.

Secondly, the company needs a prediction until the end of the calendar year about the sold quantity of added sugar and salt to have an idea about what will be the yearly result with the progress they are making. The prediction should capture the seasonality of the products.

Thirdly, they need to know the driver items of added salt and sugar quantity to make changes. Identifying these products is a key for reaching the goal, because simply improving products with high sugar and salt but with low sales or products with high sales but irrelevant sugar and salt content will not solve this problem. Furthermore, the company also wants to know the added salt and sugar amount of their products in the warehouse.

2. Analysis and prediction

To perform the project, I firstly prepared the product's details including the salt and sugar content per 100g. The relevant products of 2015, 2019 and 2020 were included in my dataset, some of them with different salt and sugar content in different years, and some product had more supplier with slightly different salt and sugar amount in them. To calculate the salt and sugar amount per products I firstly needed to get the product weight in gram. Most product had sale size variable in the data, but some product had the sale size only in the description, so I extracted it from the text using complex regular expression syntax. Then I was able to calculate the sugar and salt amount per product which I can merge with sales data.

The sales data includes every transaction on product level, so I aggregated to daily sales quantity for each product keeping in mind the weighted and not weighted product categories and the returns too. Then filtering the non-relevant products was important in order to efficiently deal with big data.

Then I merged the sales data with the product's salt and sugar content to get the sold sugar and salt quantities, which I reported on yearly level and monthly too. Then I identified the top 20 products regarding the salt and sugar amount sold. The top items show similarity with the last year's report to great extent but also there are some seasonal products considering that it is summertime. However, the actual dates are parametrized, so the program will always report the current year and the last full month.

Similarly, I performed the merging and calculation on the warehouse dataset containing the products and quantities currently are in the warehouse. I reported the full salt and sugar amount and the quantity for each product too. The top items of the warehouse were not in line with the sold products, because the warehouse had the most salt and sugar in durable products like drinks and canned products especially regarding sugar.

To perform the prediction part, I used 3 types of models: a baseline model, prophet model and light gradient boosting machine. The aim of the baseline model was just to use a simple algorithm and compare the error metric to the other models. Therefore, I applied a model which predicts a constant sales value for each product, which is the sales quantity on the last day of the train set for each product. On the 1 year-long test data and predicted the constant value for each item. I measured the error of the daily sales quantity on item level with the mean absolute error (MAE) which was 2540.

The next model I applied if a time series forecasting model developed called Prophet. The model used time series information like trend, seasonality and holidays and days before holidays, which I added specifically considering the Hungarian calendar event, because these are undoubtedly important for food sales in retail. The model performs well on one time series, however applying it on 3500 different products is not optimal. The model is not scalable and applying on one-by-one product takes much more time and computational power than other, similarly good models. Despite that, I managed to run it for every product once and it resulted 5595 mean absolute error.

My final model was an ensemble algorithm, namely light gradient boosting machine (lightgbm). Lightgbm is an efficient and open-source implementation of the gradient boosting ensemble algorithm. I used several features in the model to predict the daily sales for 365 days ahead, like lag sales, lag consumers, lag price, discounts, holidays and other calendar-based features like month, day of the week and so on. Similarly to other models, I tested it on the previous year, and it performed 1438 MAE. The huge advantage of this model is that it can classify the products so can make prediction for each item with one training. Besides, it can also capture the relevant external factors of sales like discounts which is not incorporated into the prophet model but had huge importance.

3. Summary

Therefore, I used lightgbm as my final model and trained the model on the time series until the end of the last month and made prediction until the end of the year. I applied it with parameters, so in the next month the training interval will increase, and the prediction interval will decrease since it's always until the end of the current year. Finally, I reported the products with the highest salt and sugar amount based on the prediction and added to a dashboard with all the other relevant information needed, including the ratio of salt and sugar compared to the weight of the products.

This project enabled me to perform a full data science workflow from understanding the business needs, through data cleaning, feature engineering, modelling and presentation of the results and to gain experience in technologies I have never worked with, like Azure Databricks, and Pyspark and. The code is written mostly in Python and using some SQL which made me able to gain routine and confidence in these programming languages too.