Summary

Egis Pharmaceuticals on its digital transformation starting journey is committed to data driven optimization projects. This commitment set on its way the production optimization project related to one of its main product.

The goal of the project was to identify the key factors responsible for the low outputs of the production units and by understanding these parameters decrease the number of low output production units, thus increasing the mean of overall production. As this product is a very expensive product, increasing the mean of production units would on one hand increase the profitability of the product and on the other hand solve issues related to the production and sales planning of the product.

The project was performed by a small team under agile methodologies. The members of the team were selected prior the kickoff of the project and were appointed to an agile and scrum training. The product owner was supported by an appointed (third party) product owner mentor throughout the project. The involved stakeholders were also appointed to different agile and scrum trainings before launching the project.

The members of the team involved cross functional persons who were experts in their own area of expertise.

The members of the team:

- Project owner: unit lead at the Agent Production Facility
- team member: chemist production professional (technician), team lead at the Agent Production Facility
- team member: data engineer from the Automatization and Maintenance Department
- team member: data scientist, the data scientist of Egis
- team intern: intern with temporary contract
- team scrum master: the scrum master from a third party company taking care that the agile methodology and scrum framework is being utilized and facilitating the team's work

The project was performed in 4 sprints using scrum framework. As it was mentioned previously the aim of the project was to identify the most important drivers of production output ratio with the help of data and machine learning tools.

The project involved data collection and data cleaning, data joining, exploratory data analysis (correlations, distributions, descriptive statistics) and linear regression and tree based machine learning models. In the early session of the project the data engineer was reliable to identify the relevant data sources and databases. As a next step he was reliable to collect the relevant data from different data sources. Meanwhile the intern was helping the team by inputting the data from the production sheets into excel files. As this was a manual task the team took extra caution in double-checking the input values.

When all the relevant data was collected there was a plausibility and validity check performed on all data sources. As a next step the data scientist joined the databases and the data exploration and analysis could start.

The analysis was performed both by the data scientist and the chemist team member. As there were more than 25 parameters involved in the production process as first step the team defined 4 stages of the production. As a next step there was a correlation and distribution analysis, descriptive statistics report and modelling performed on all – the previously selected – 4 stages. The performances of the models varied, there were better and poorer performing models. As for machine learning models, I used both linear regression and tree based models (random forest and gradient boosting models) in order to find non-linear relationships as well. Mostly the gradient boosting models seemed to perform better. As for the model score I used both mean absolute error and root mean squared error. For better interpretability and in order for the stakeholders to understand easier, for the sprint demos I used mean absolute error as a metric for the models.

After learning the performance of the models for all 4 stages, taking in consideration the results of the correlation analysis and model coefficients and feature importance values, the team selected the most powerful and important parameters. Then by using these parameters we created the final model.

Although the final model performed better than all the models before (on 4 stages), the team was not satisfied with the results as the score of the model was not considerably better and also the coefficients and feature importance values were not straightforward.

As a next step and turning point I decided to split the output ratio based on output ratio values:

- higher output ratio: these were the best production units
- average output ratio: these were the production units with average values
- low output ratio: these were the poorly performing production units

After splitting the data in the above mentioned categories we created models on all categories and found out that there is a significant difference on the model performances based on the categories.

The best performing model belonged to the higher output ratio, the second best performing model to the average ration and the worst performing model belong to the low output ration production units. These differences made it obvious to the team that there must be a parameter which is not observed or measured by us (i.e. we do not have any data about it) which must be causing the low outputs. This unknown parameter – most probably – could be a pollutant which is present in higher amounts in case of low output ratio units, slightly present for average output ratio units and not at all present in case of high output ratio units. That is why the model performances are acting so differently for the three categories.

All in all by the end of the project the team was not able to clearly identify the parameter that is causing the low output ratio production units (based on the findings there is a high chance that we are not measuring/monitoring that parameter) but found some key production practices which should be changed for higher production units.

Our suggestions were accepted and tested on a few production units so far, and they seem to work, thus increasing the production ratio mean of the product.