BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case 4: Services Demand Forecasting

Palm & Company

Pedro Santos (M20190420)

Ana Cláudia Alferes (M20190932)

Lennart Dangers (M20190251)

Michael Machatschek (M20190054)

May 2020

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. INTRODUCTION

Since the liberalization in aviation in Europe, a wide range of low-cost airlines entered the flight market. This brought more opportunities for tourists and accordingly to this, businesses, such as transfer services, are also growing. Due to technical progress and the use of mobile apps, bookings are significantly handier these days. Where more opportunities and flexibility suit the tourist, it makes accurate demand planning more difficult for companies in that business field.

This report shows our approach to create a demand prediction model for the company Yellow Fish Transfer.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND AND CURRENT SITUATION

Yellow Fish Travel Agency operates a transfer service in the Algarve, Portugal, and their services are mainly from Faro airport to the hotels or doing the transfer for golf tours. They provide up to 500 services a day with around 22,000 passengers on average per month. It is essential to mention that they have no stop in bookings and in case their capacity is reached, Yellow Fish will subcontract the service to partners respectively competitors.

Due to seasonality and volatility, Yellow Fish has less long-term employees and must hire drivers regarding the demand. The accurate demand forecast is vital because the hiring process is cumbersome and needs around 8 weeks. This comes with a cost in money and time. The current state is often a tradeoff between too many drivers respectively cars and overbooking their capacity. Or in more monetary words, high fixed costs against costs for subcontracting and the loss of the quality control and future revenue if customers would book with the competitors again.

2.2. BUSINESS OBJECTIVES & SUCCESS CRITERIA

<u>Business Objectives</u>
The major objective is to predict the number of services on a weekly basis. The desired output is a forecast of the demand for eight weeks in advance in four weeks batches. Through this prediction, the Human Resources Department can do the hiring process in time. Furthermore, the forecast can work as an input for the already existing software to obtain a more accurate allocation of the drivers. Regarding a more monetary perspective, a precise forecast avoids costs that will occur when overestimating or underestimating the number of drivers.

As a side effect, such an analysis should deliver further patterns, which are valuable for the whole business. Here it is essential to deal with questions, such as “How high is the cancellation rate?” and “Which is the best working month regarding the number of services?”. Eventually, it is essential to deploy express the findings understandable for all stakeholders and include it in the daily business processes.

<u>Business Success Criteria</u>
From a business point of view, the prediction should have a deviation of the true demand of 10 % at maximum below and above the actual number. A further business success criterion is to obtain useful insights that help the business to understand and improve its services.

2.3. SITUATION ASSESSMENT

Resources

The project team is composed of four master students in data science and advanced analytics with different professional backgrounds. The given dataset which is described in data understanding acts as the main input for the prediction model. The main business knowledge is gained through the case introduction and the queries we had during the project. The main technology used is Python and its corresponding libraries such as Pandas, Plotly, Keras and Scikit-learn.

Requirements, assumptions, and constraints

The main constraint within this work is, with a limitation of three weeks, the time. It can be challenging to obtain answers to fundamental business understanding questions. For this reason, a successful project requires one business expert from the customer who acts as a vital link between the project team and the business itself. Since the desired outcome is a weekly forecast, the quantity of data is another constraint. While working on the project, no further assumptions beyond the information in the case introduction were made.

Risks

The quality of a prediction model is based on its input data. Both quantity and quality are essential. Even an advanced model cannot prevent unforeseen events such as pandemics or natural disasters. For this reason, the highest risk is a prediction not being within the confidence interval of 10 %. However, the model will be evaluated in detail in a section below.

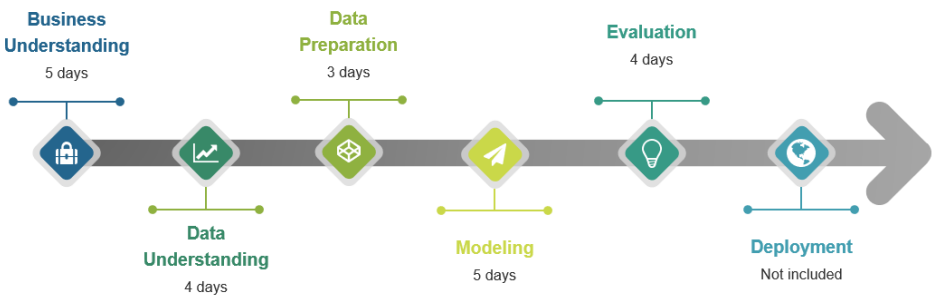
2.4. DATA MINING GOALS

Alongside the business objectives, the data mining goals are a rather technical target to reach the overall goal. Firstly, we intend to do the necessary preprocessing to be able to do proper analysis even in future tasks. Secondly, we will perform a detailed data exploratory analysis to get valuable insights and detect patterns. Eventually, the final goal is to apply the best working algorithms and create an accurate prediction model to do the desired demand forecast and additionally implement it in the existing processes.

Besides other metrics, the main metric is the mean absolute percentage error (MAPE) which measures the deviation between the prediction and the actual value. In this project, the desired MAPE should be below 10 % that is also the main data mining success criteria.

2.5. PROJECT PLAN

This project follows the CRISP-DM process. The most important characteristic that stands out within this process model is, that it is not hierarchical, and one can move one step back.



Please note that the time in each process is a rough approximation. During all these phases several interactions between the phases can be done to improve the decisions in previous steps. Therefore, and due to dependencies among each step, an agile adjustment of the project plan is possible.

3. DEMAND FORECAST PROCESS

3.1. DATA UNDERSTANDING

The goal of the data understanding section is to explore and explain the data based on statistical methods and the obtained knowledge in business understanding. Following the CRISP-DM process, this part contains a data collection and description report as well as a data exploration.

3.1.1. DATA COLLECTION AND DESCRIPTION REPORT

Data Collection Report

The examined dataset contains data about a Portuguese Transfers company known as "Yellow Fish Transfers". The data contains information about the years 2016, 2017, 2018 and 2019 through different CSV files. The variables display information about the location and dates of the services as well as the date of the bookings, type of service, information about the driver and vehicle, the status of the booking and service. For some services we also have information about the supplier, meaning that YFT did not have enough resources to make this service. Furthermore, there is feedback file, containing the evaluation of the service made by the customers.

Data Description Report

The main dataset (Manifesto) has 24 columns, which represent different features. The number of rows is 370,264 and each row represents one service. Most of the features are categorical variables containing strings or are dummy variables. A detailed description about the datasets can be found in the notebooks.

3.1.2. DATA EXPLORATION AND QUALITY REPORT

The data exploration process was initially made for all the different CSV files, in order to have an overview about the data. In a more advanced stage of the EDA, we proceeded to the aggregation of the data to a weekly level in order to have more accurate representation of our targeted forecast dimensions.

Data Exploration Report

Looking into the plots made in the Notebook it was possible to reach some interesting conclusions. In the distribution of the number of services through weeks, we can see that the weeks that have the most services are between week 20-35, corresponding to May, June, July and August. This "summer trend" is also visible in other features, like the number of drivers and average lead time, since the company has more market demand, the customers usually make their booking more in advance. In some cases, as referred before, the company did not have enough resources to complete all the services and had to give up services to their competitors. Once again, this happened with more frequency on the high demand season, but in this case, we have some clear peaks in the week 22 and between week 36, 37.

Contrary to the other variables, the cancelation rate reaches its maximum in December with 18.2 % cancelled services, whereas in the rest of the year there is cancellation rate between 11 % and 14 %. Concerning the average rating, we can see that through the year it does not show many changes, presenting most of the time ratings between 3.6 and 3.7. However, a decrease of the ratings is visible in the end and in the beginning of the year. In a last note, the data also showed that around 60 % of the clients are English, presenting this way a majority comparing to other nationalities.

Data Quality Report

In the EDA process there were found some minor problems with the data. Starting with the Feedback file where some data incompatibilities were found, i.e. in some cases the rating of the service was too low, but the comment attached to it was positive, making us believe that maybe some customers did not get the scale of the rating right.

In the Manifesto file we found some inconsistencies as well. In the features Booking Status and Service Status, since the Service Status is depending on the Booking Status, once the Booking is marked as "cancelled" or "no show", the Service should have the same status. In some other features like Flight Number and the GPS Coordinates were found some problems like a high number of null values and incorrect information, so we decided to not use these variables. In an overall look we can say despite these minor problems that the data has a good quality to obtain reliable results.

3.2. DATA PREPARATION

The goal of the data preparation phase was to obtain one final dataset with several, possibly relevant features, aggregated on a weekly level, that initially can be used as a starting point in the modeling stage and eventually builds the basis for the final feature set. Our plan was to include rather more features right from the beginning to have several options to experiment with and reducing the features step by step. Eventually, our final feature set was only a small fraction of the initial dataset.

Data Selection

The following table shows the features that we used, how we integrated them in our final dataset, and the rationale for inclusion.

Feature Name	File	Integrated on weekly level as:	Rationale for including
supplierID	M	Proportion of outsourced services	It can be seen as the error of the previous prediction
whichway	M	Proportion of arrivals	Many arrivals in one week can lead to more departures in the following
paymentType	M	Proportion of cash payer	Possible relationship between the type of payment and cancellation risk
customerID	M	Proportion of booking with returned customer	Development of customer retention rates can influence future demand
typeOfService	M	Proportion of oneway trips	An increasing number of two-way can lead to higher demand
serviceDate, bookingDate	M	Number of services already booked for the forecast period	Demand already on the books can be a indicator for the actual demand
		LeadTime	Lead time in combination with bookings on the books can be a good indicator for the demand
pickupGPSCoordinates, dropoffGPSCoordinates	M	Average driving time	Development of driving time can influence demand
driverID	M	Unique number of drivers	Experimental feature
adults, children, babies	M	Average number of persons per service	Bigger groups may need more cars

bookingStatus, serviceStatus	M	Proportions of "done", “canceled”, “no show” services	Development of cancellation and no show rate influences actual demand
		Number of services per week	Demand of last weeks as indicator for future demand
All rating columns	F	Average rating per week	Customer satisfaction can influence future demand
Pais_id	Cu	Proportion of customers from Portugal, Ireland, England	Experimental feature
newsletter	Cu	Proportion of newsletter subscriber	Related to customer retention
loyaltyCard	Cu	Average points on loyalty card	Related to customer retention
cancelDate, serviceDate	Ca	Average cancelation time before actual service	Cancelation time in combination with bookings on the books can be a good indicator for the demand

M: YFT_Manfiesto.csv | **F:** YFT_Feedback.csv | **Cu:** YFT_Customers.csv | **Ca:** YFT_Cancellation.csv

We made use of most of the available variables. There were some variables that we did not use. The reason for exclusion was either quality issues or because we did not see any relevant information that can contribute to our solution.

Data Cleaning Report

As mentioned in the Data Quality Report, the data had overall good quality. There were some minor inconsistencies that we fixed before starting the modeling. Many of the issues were related to a problem with inconsistent booking and service status. After consulting our contact at YFT we managed to resolve all the issues related to this problem. Detailed documentation of each data cleaning step can be found in the ETL notebook.

Feature Engineering (derived attributes)

In total, we created several new features such as the average driving time, the average rating per week, or the average time between cancellation and service date. However, in the end, it turned out that these features are indeed insightful but did not improve our final solution. In the following paragraphs, there will be more details about two derived features, that had a more important impact on the result.

Positive and negative outliers

The goal of our proposed solution is not a full automation of the process. It is about getting the best prediction as possible. Keeping a human in the loop, so giving a human an effective way to influence the prediction of the model, allows you to incorporate subject-matter experts’ opinions in the forecast. For this reason, we created a dummy variable for positive and negative anomalies. After feeding the model with several instances of positive and negative outliers, it effectively reacts to those inputs and adapts the prediction accordingly. Now this variable can be changed manually, if some extraordinary events will occur in the forecast period.

Lag features

After selecting the features and creating new ones, we needed to create special time-dependent features that can be used in a time series forecasting with machine learning models. These features are also called lag features and they contain data from previous time steps. Having data points from previous time steps as an input, we were able to feed our models with information about seasonality and short- and long-term trends in the data. We experimented with several combinations of lags and eventually, we choose the ones that can be seen in the table below.

Final Dataset

Our final data set used for the final model training and evaluation includes the following features:

Feature	Lags
numberOfServices	12, 13, 14, 51, 52 weeks ago
pos_outlier, neg_outlier	1, 2, 52 weeks ago
avgLeadTime	52 weeks ago
cancelRatio	52 weeks ago
weekOfYear	-

3.3. MODELING

Our modeling process has four stages. The first stage consists of model selection, by testing several models against each other, with a predefined set of features, in order to evaluate which performs better. For this purpose, we have selected eight different models (linear models, ensemble tree models as well as neural networks) and compared their results on the metrics Mean Absolute Error ("MAE"), which translates to the average absolute difference from the prediction to the actual value, and Mean Absolute Percentage Error ("MAPE"), which is the percentage deviance of the same error to the true value. The first feature set contained only 12-week and farther lag features as well as some time indicative features, like week of year. Here, the fully connected dense Neural Network performed the best and, therefore, was the model selected and used for the following stages.

The second stage consisted of developing a Walk Forward approach to validate our model. This approach differs from the previous in many ways: we use the latest data available to train the model and use it to predict the following week, which then is used as an input to predict the week after and so on and so forth, until the target weeks have been predicted. Although this ensures that the model always uses the most recent data, it failed to predict accurately a large window of weeks, as it is the goal of the project. Nevertheless, we believe that if the prediction period were rescaled to a shorter window, it would have been possible to retrieve results well within the 10% MAPE range.

The third stage was applying a Rolling Window approach where we defined several window sizes and its moving average and uses the latter to make prediction. The application of the Rolling Window with the moving average was not successful and, therefore, discarded.

Finally, the fourth stage of modeling consisted in selecting the best working approach, tune it and improve feature selection. Our final model is composed by 16 features, including 12-week or later lag information as well as expected anomaly information. This combination of features allows to make prediction with an average error of 12,5 %, which, even though higher than the error interval requested, represents a great result due to the unprecedent factors in the test period, such as the bankruptcy of Thomas Cook airlines.

It is important to note that the process was iterative, since whenever considerable changes were made – be it to the ETL or feature selection and engineering – the process would restart and every stage and model evaluation was redone to ensure the best results.

As the model explainability was not a concern for the project, there were no further analysis made on the regressors' coefficients nor any additional exploration on the interpretability of the deep net.

4. EVALUATION AND FUTURE IMPROVEMENTS

One of the main constraints within this project was the time and the limited availability of data. In terms of business objectives, we could obtain valuable insights such as the average lead time or a cancellation ratio per week. Furthermore, a prediction model was created and is ready to be implemented in the day-to-day business.

The model has a performance slightly lower than requested, but as it consists of a prototype, we are confident that, given more time, a better result could be achieved. Nonetheless, the average error of 12.5 % also represents a positive outcome taking into consideration the volatility of the year 2019, due to the outlier events already mentioned. In retrospect, in May 2019 the model underestimated the demand the most, which can be connected to consequences of the Brexit withdrawal agreement that was postponed in late April 2019.

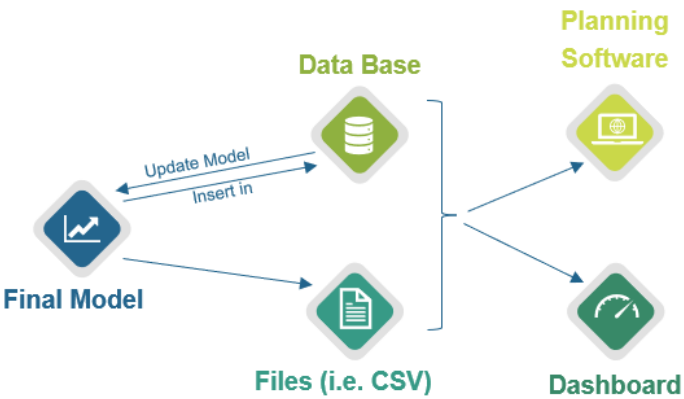
For a future version, we would focus on the development and improvement of the walk forward model, as it is able to use more recent data in the training as well as return a better validation of the model's performance. Business wise, we also believe that a walk forward approach could be more suitable, but more testing would be needed.

Finally, the model would also improve with the addition of more data and a more detailed optimization/tuning. All these aspects will be covered in the case of a future implementation.

5. DEPLOYMENT AND MAINTENANCE PLANS

Deployment

The last step of the CRISP-DM process is slightly more independent from the other steps. Once the final model is evaluated, the deployment phase takes the results and provides a strategy for deployment. In this case, several deployment outputs are thinkable and depend on the customer's wishes and infrastructure. The following figure displays the possible deployment steps in a nutshell.



One option is to build an ETL-process to load the predictions into the Yellow Fish' database. This process should be executed with automated jobs. ETL pipelines can be built with python or open-source software, such as Pentaho Data Integration. Furthermore, new business data, that is collected through new services and stored in the database, should also update the model. With this approach improvement of the model, due to a higher quantity of data, can be obtained. Besides that, another possibility is to extract just single files, for instance, a CSV-file. However, we would recommend implementing an automated ETL process to assure data consistency. In both approaches, the output can be used to feed data into the already existing planning software of Yellow Fish. Additionally, a dashboard should be created to provide the management with live forecasts and crucial business figures. For this reason, software, such as Microsoft PowerBI, Tableau, or Qlik can be used to create and share dashboards with the management.

Maintenance

The goal of maintenance strategy is “to avoid unnecessarily long periods of incorrect usage of data mining results.” (Chapman, P, 2000). Especially after implementing the model in the day-to-day business, monitoring and maintenance gets crucial. To maintain the model, it is recommended to keep training it before any prediction period, since having the most recent data available is necessary for the time series prediciton. Furthermore, a frequent validation cycle in the first deployment phase is highly advisable. Since this model is not proofed by new data, this step is necessary to avoid possible inaccuracy even though the model is evaluated.

6. CONCLUSIONS

Particularly in uncertain times like these, companies need to adapt their strategic and operational plans wisely. Data can offer additional intelligence for planning capability.
We think that the combination of computational and human intelligence can offer the most precise prediction of future scenarios. Our solution offers you enough flexibility to react to short-term events that affect your business. Eventually, it lets you save costs and build your customer base by balancing over- and underestimation of the demand.

We want to emphasize once more that we proposed you a prototype with this solution, that can be the starting point for a more advanced solution. With more training data and further model optimization, there is still a lot potential for improvement. We are dedicated to realizing this potential for you.

7. REFERENCES

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Cross Industry Standard Process for Data Mining. In *CRISP-DM Consortium*.