BUSINESS CASES WITH DATA SCIENCE

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

Hotel 2

A Cancelation Prediction Tale

Group T

Ana Amaro, number: m20200598

Filipe Lourenço, number: r20170799

Gonçalo Almeida, number: m20200594

Guilherme Neves, number: r20170749

March, 2021

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

INDEX

1.Business Understanding

1.1. Introduction and Presentation of the Business Objectives

Recent changes in the hotel industry, mainly caused by the emergence of Online Travel Agencies, created the need for hotels to find a way to deal with excessive bookings' cancellations. Like so many other hotel's owners, Michael tried different options, like aggressive overbooking, to attempt to solve this problem, but without luck, so he decided to contact our company to use a data-driven approach. The focus of the work is, at this stage, exclusively over Hotel 2, a city hotel from the chain Hotel C.

The primary business requests from Michael were for us to develop a predictive algorithm that would allow us to predict net demand for the hotel (for him to be able to implement an appropriate strategy of overbooking) and identify the most likely customers to cancel the stay (which can allow for them to be offered discounts/incentives to discourage them from doing it). This problem is fairly complex as an excessive overbooking policy leads to costs and reputation damages, while the opposite may lead to excessive inventory that cannot be sold at the expected price (or at all).

Based on the goals Michael expects to achieve, the business success criteria can be split into two. The model should be able to identify properly the customers more likely to cancel, in such a way that Michael can implement a strategy of incentives to reduce the number of cancellations from 42% to 20%, while also providing the tools to significantly improve the demand forecast process, allowing for an efficient overbooking policy.

1.2. Situation Assessment

Michael provided us with a CSV file containing records of the bookings for Hotel 2 from the 1st of July 2015 to the 31st of August 2017. Inside we had 31 variables referring to characteristics of the booking and customers, including the result of the booking (either cancelled or not). A significant part of the variables was anonymized and pre-processed before reaching our team and has been described in an associated metadata document.

We agreed with the timeframe purposed of one week to prepare a presentation of 5 minutes to the management team and to deliver a full report regarding the steps we followed and the recommendations we want to provide. Moreover, the main costs agreed are the working hours our team will devote to the project and that will be billed to Hotel C in function of the benefits achieved, that are closely related with the objectives already mentioned. The main risks associated with this task are very generic and arise from a potential lack of representativeness or quality of the data we obtained, which can easily be mitigated by asking for more data if along the way we find reasons to suspect this happening. Lastly, from a terminology point of view, we need to understand the concept of Net Demand in the hotel industry, which refers to the total reservations minus cancellated bookings.

1.3. Data Mining Goals

The data mining goal of this project is to construct a classification algorithm that predicts, for each customer passed to the model if he/she will cancel the booking or not, keeping in mind the costs

associated with each error made. Moreover, based on how the problem was presented and to also give more resources to the management and marketing teams to in the future explore different alternatives, we intend the creation of a white-box model, which will also help in the task of reducing the cancelation rate as the marketing team will be able to know which customers to target (e.g., a specific distribution channel or agent).

In order to measure success, we will use mainly the F1 score to assess the quality of the model, and although the definition of a good value for F1 score varies from problem to problem, we hope to achieve at least a value of 0.75. Instead of an F1 score, we could use Accuracy, as the dataset is fairly balanced, but considering the overview Michael provided us about his business we decided it would be better to evaluate the model based on a metric that takes into consideration the costs of being wrong in both directions (instead of one that just considers the number of right predictions), given that, in this context, a false negative can lead to excess inventory and all the costs associated, while a false positive can lead to a customer being offered a discount/incentive which he did not need and can also conduct to overbooking. Furthermore, considering we need to provide Michael with the most likely customers to cancel, we believe it would be important to achieve a lift of, at least, 2 (in a point that allows the reduction from 42% to 20% of cancelations), meaning that we should offer Michael a way to contact customers in such a way that for a certain percentage of the total bookings he contacts, he will be reaching at least twice that percentage of the bookings that would be eventually cancelled.

1.4. Project Plan

Concisely, our project will be based on the following steps: 1. Retrieve the dataset, gather some superficial understanding of the data and the variables and try to spot easy to detect problems, by using techniques like statistical analysis of the variables; 2. Deeper exploratory analysis and insight extraction about the data; 3. Clean the data by acting upon the insight found at step 2 (e.g., remove outliers, solve missing values and other problematic values); 4. A small review of the steps done until here, by confirming if all the problems encountered were solved; 5. Define the model(s) to use, create the conditions to apply it (e.g., split train and test set) and construct it, based on the data analysis done and on the business and data mining goals (which as mentioned already implies a white box model); 6. Attempt to improve the model(s) performance by tunning the parameters; 7. Verify if the model meets the data mining success criteria, and interpret the results achieved in light of the business context; 8. Present the report to the management team and if the approval is met, deploy the plan.

2. Data Understanding

The data collection was straightforward, as we only had to import a single file and that already suffered an initial pre-processing step.

While analysing the data with different techniques (statistical descriptions, boxplots, histograms, pair-plots, correlation matrixes and count plots for the categorical data), we detected some concerning situations. Firstly, there seemed to exist significant outliers, categorical variables with high cardinality, some missing values (that according to the information provided should not exist) and duplicated entries (that was confirmed with Michael as being always different customers). Furthermore, there seems to also exist some incoherencies, or at least not relevant entries, like bookings with no nights spent at the hotel and/or with no children, adult or baby associated to a room (in some of those scenarios, we acknowledge that there are customers that only book services like spa

and/or restaurant, which is not interesting for our purpose, so these bookings do not constitute "quality data"). On this note, we also found that some of the customers were considered a repeated guest, although never having cancelled or not cancelled a previous booking, while the opposite also happened, which in both cases constitutes an impossibility that we will have to consider as an error in the data collection process (or in the imputation of missing values when this dataset was previously pre-processed). Lastly, we also detected some problems with the extraction of the variable DepositType, which seems to have been wrongly collected.

3. Data Preparation

Before attempting to address some of the problems presented in the last section, we decided to drop some of our data. Firstly, we decided to drop the variables ReservationStatus, ReservationStatusDate, AssignedRoomType, BookingChanges and Country as those do not help in the predictive model construction, as they only become fully known in a phase where it is already clear if the customer cancelled or not the booking (in the case of the BookingChanges, we considered that any change will be made after the booking, contrary to, for instance, TotalOfSpecialRequests, that we assume was collected as it was on the moment of the booking). Note that in the particular case of the ReservationStatus, we decided not to use the "No-Show" as the third label to our problem because, although this constitutes a third type of customer that pays the stay (like the ones that do not cancel the stay), but allow for overbooking (like the ones that cancelled), the proportion of customers of this type in our dataset is not significant (mainly considering the next steps will make this segment decrease even further), and for the predictive model, categorizing these customers as "Cancelled" gives Michael a more clear view of the overbooking that can be done. Additionally, we also removed the variable DepositType due to the lack of quality mentioned above. Lastly, we also decided to not keep the variable ArrivalDateWeekNumber (that presented a very significant Phik correlation with ArrivalDateMonth) and ArrivalDateYear (as this variable represents non-repeatable events, which in turn makes it not useful for this problem). Furthermore, we decided to keep variables as the month of the arrival as we believe there is forecast power associated with those (e.g., in a specific month there is more or less likelihood of cancelling). Note that we decided not to act on the duplicated entries, as we were assured they were different customers and considering the data collected corresponds to what we believe to be a representative sample (considering it covers all customers from a relatively large temporal range), we want our model to be able to face a dataset as close to a real-world scenario as possible.

Regarding the inconsistencies found, we decided to remove those entries, as in some cases it represented data that is not interesting for our problem (e.g., customers who did not book a room, but a hotel service), and in others, it represented what it seemed like problems of data collection (e.g., repeated customers that did not show up in the records as having cancelled or not a previous stay).

The next step we decided to take was to remove some of the most significant outliers from our dataset, and in this regard, our team used a two-step confirmation for outlier detection (only considering a customer as an outlier if he was both an outlier in the univariate and multivariate space) based on the IQR method and with the Mahalanobis distance, that we decided to set with the value 2 for the multiplier and 5% for the most extreme level to accept, respectively, which ended up removing 2.17% of our entries (considering as baseline the dataset with the already mentioned steps).

Moving on to the construction of data, here we decided to solve some high cardinality situations in the categorical data. In the case of the variables Company and Agent, we only kept labels that had more than a thousand customers belonging to that segment. This led to Company being left with only two options (being a company's booking or not) and Agent with 12 labels (including the label indicating that there was no agent and a created label compromising all the other cases). The variable "ArrivalDateDayOfMonth" was also deconstructed to represent the week of the month where people expect to arrive, as in the hotel industry this is usually a more relevant metric. Additionally, we also changed the metric nature of some features that although numeric had a very specific distribution of values: both the variable Babies and RequiredCarParkingSpaces presented a significant amount of 0's in relation with the remaining values it could take, so we decided to binarize this variable in either requiring a car parking space/bringing a baby or not. After this, we confirmed visually that the steps undertaken had improved the problems found in the Data Understanding phase of this report. Lastly, and before proceeding to the modelling phase, we used an encoder to convert the categorical variables into binaries.

4.Modelling

4.1. Modelling Technique and Methodology

As previously mentioned before, given our desire to provide Hotel 2 with a mechanism that can be applied to different contexts, like marketing efforts (considering the Online Travel Agencies' impact in the digitalization of the sector, targeting the right customers can be even more crucial), we decided to develop a white-box model. In this particular case, we believe that a Decision Tree may be the most indicated, as it is decomposable in logical steps that Michael and his team can follow. Additionally, this algorithm also allows us to retrieve the most important features according to the metric used (e.g., entropy), which can prove to be useful in the business context.

The Decision Tree is a non-parametric model so it does not make any assumption about the data distribution, and we will not need any processing steps (e.g., standardization) other than the ones already taken (e.g., the encoding of the categorical features). Nevertheless, it also has some disadvantages, being especially prone to overfitting and highly unstable when compared with more robust models, thus creating the need for a careful parameter tunning phase.

Regarding the experimental methodology, we decided to use a simple stratified split, with around 20% of the data being reserved for the testing, and the remaining for the training phase. Although this step was done at this moment to comply with the CRISP-DM methodology, from a pure data mining/machine learning perspective, this should have been done before the majority of the analysis made so far (e.g., outlier removal), to avoid data leakage. We will use the test data to verify if the data mining goal was reached by tracking the metrics mentioned before. Moreover, under the need to do parameter tunning, we will use a cross-validation mechanism built with 5 stratified folds (to assure the representativeness of each fold) to assess the impact of the different parameters on the model quality, but as mentioned, the final evaluation will rely on the original train and test split. Additionally, and considering the situation related to the "duplicated" entries, and although the test set will be left intact (to assure the model will be tested in a close to "real-world" scenario), we will attempt to use a training set including duplicated and other excluding those. Finally, to assure reproducibility and comparability, we will rely on a random state definer and it is also important to

remark that our target variable will be IsCanceled, a binary variable (creating thus a classification problem).

4.2. Model Construction and Assessment

Regarding the construction of our model and considering the mentioned use of a cross-validation method for parameter tuning, we will start by using a Decision Tree with all the Sklearn's implementation standard parameters, employing a GridSearch attempt to find the parameters that give the best F1 score. For this particular problem, the parameters we tried to tune were the criteria for the split, the minimum number of samples to constitute a leaf, the maximum depth of the tree, the minimum number of samples needed to make a split, the maximum number of features to use per split, the class weights and the parameter for cost complexity pruning.

From all the combinations attempted, the parameters that gave the best F1 score were: using entropy, a balanced class' weight, considering 3 the minimum number of samples needed to constitute a leaf and a cost complexity pruning penalizer of 0.0001. With this model, we were able to achieve an F1 score of 0.8. Moreover, the best result was achieved with the "duplicate" entries in the train dataset (which would probably not be the case if the training set was overfitting due to some specific bookings being repeated).

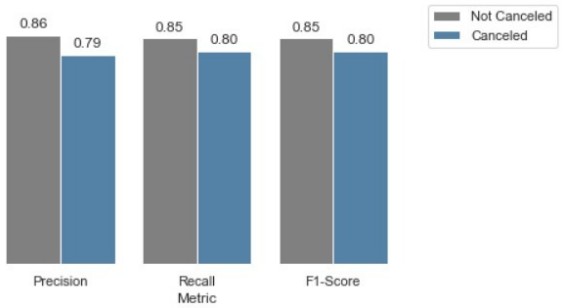

Figure 1: Precision, Recall and F1-Score's values for the test set.

From a pure data mining evaluation of the model, those were the main results to present, and we consider having achieved the proposed data mining goals, achieving an F1 score above the proposed 0.75 and a lift of at least 2 until around 38% of the total dataset is contacted. Furthermore, even from an accuracy point of view, that was not our main metric of evaluation, we achieved a result of 0.83, which implies that we can predict correctly 83% of the bookings' result at the time the booking is made.

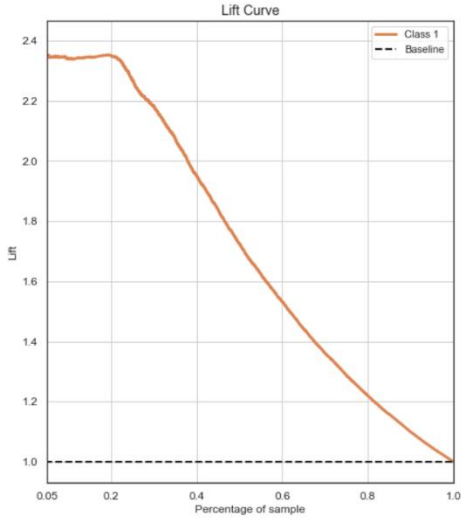

Figure 2: Lift Curve.

In the next section, we will explain how our model can be translated in the context of the hotel, particularly how the manager can leverage it to decide the appropriated number and type of customers to contact to try to persuade them not to cancel their stay and how Michael can apply it to forecast demand according to the business needs.

5.Evaluation

5.1. Evaluation Results

From the business side, the success criteria will take longer to be achieved, as they will require "in the field" verification. Despite this, we believe to have created a model that satisfies the business goals and that will meet the expectations when deployed.

The first result we would like to point out is how the model provides the most important features to predict if a customer will cancel the reservation or not, which can be extremely useful in future marketing efforts from Hotel 2, as mentioned before, which may also contribute to reducing the cancellation rate (particularly knowing the most important agents, customer segments and market segments).

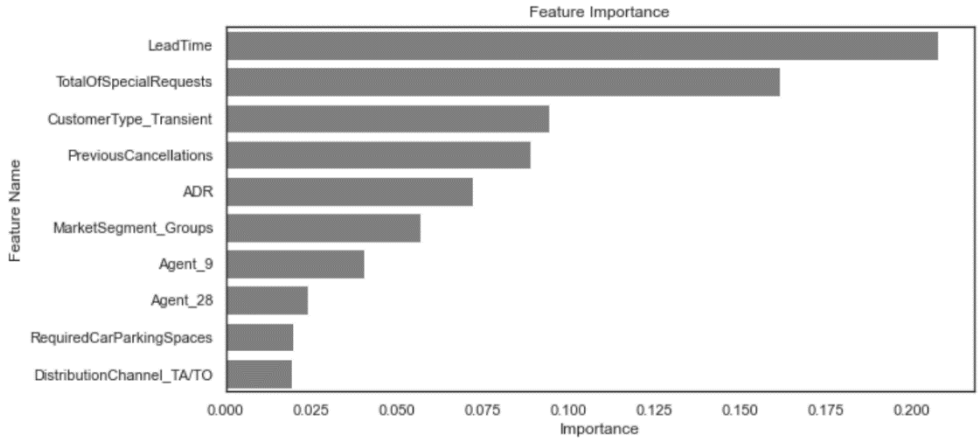

Figure 3: Feature Importance for the top 10 variables.

Furthermore, and now directed to the business objective of creating a forecasting model, we believe that the solution created will allow Michael to be able to more efficiently and effectively define an overbooking strategy, as requested. In this regard, we provided a tool that considers that both overbooking and having spare inventory is expensive, which ended up being able not only to predict 80% of the true cancelations (recall), but also being right about each cancelation it predicts 79% of the time (precision). On this same note, the model ended up producing around the same amount of both types of errors (1282 false negatives to 1342 false positives), which according to our understanding is a good result due to the high costs of having a strategy that creates too much of each risk, as in previous endeavours conducted by Hotel 2.

		Predicted	
		Not Cancelled	Cancelled
True	Not Cancelled	7637	1342
	Cancelled	1282	5152

Table 1: Confusion Matrix for the test set.

In what concerns identifying the customers that should be contacted to persuade them to stay, our recommendation would be to contact all customers that have a higher probability of canceling than 83.6%. This result was achieved using the concept of lift, where our model suggests that this threshold would allow us to contact around 57% of our customers that are going to cancel the stay, while only contacting 25% of our total customer base. Although this result will need further discussion with the management team - as there is the need to evaluate the economic costs of offering a discount/incentive to someone who does not need it (false positives) in comparison with the drawback of not being able to contact as many customers that will cancel as we would like, by improper classifying them (false negative) -, we believe this threshold would allow for an efficient approach, being in line with the Hotel 2's desire to lower the cancelations from close to 42% to 20%.

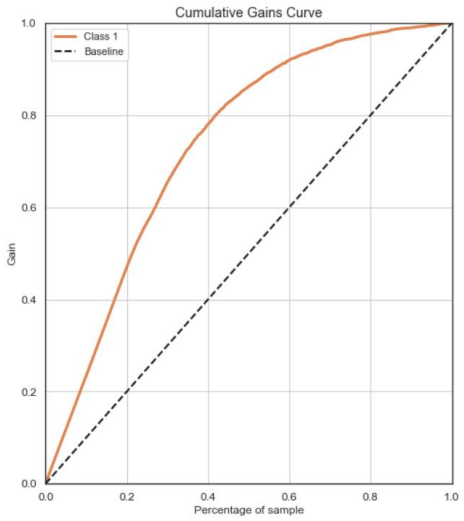

Figure 4: Cumulative Gains Curve.

Overall, we believe to have developed a well-constructed model that is not only able to be applied to this forecasting problem, but that will also help the marketing department in the task of finding customers less likely to cancel, which will also help the hotel to dilute the cancelation rate from 42% to 20%.

Regarding future data collection efforts, we advise each customer to have an identification ID, so there are no doubts about the uniqueness of the entries, and that all variables should have a timestamp associated so that there would be a more precise notion of until when (before the arrival of the customer) each variable was updated. Still regarding data collection, we believe that more variables could be collected and some data problems could be fixed (e.g., the poor construction of the variable DepositType).

Lastly, and considering there is the intention to contact the customers more likely to cancel, we believe the model proposed to forecast demand may start to not be as efficient as it does not consider this, so a new model could be developed, after an initial data collection phase, to identify the most likely customers to revert a cancelation, so the predicted net demand can follow a new formula: "total bookings" – "predicted cancellation" + "predicted reconversions".

5.2. Review Process

When reviewing the data mining process, our team did not spot significant mistakes or failures. The process followed the structure initially proposed in section 1.4. of the report and complies with the CRISP-DM methodology. Being this said, from a more technical perspective, we again reinforce that the process of creating a train and test set should have been done earlier than suggested by the CRISP-DM. In this particular dataset, we had a challenge with the "duplicated customers", but considering the reasons already mentioned, we believe to have followed the correct procedure, according to which we even attempted to train our model without those entries. Obviously, we may risk having some overconfidence in our metrics (some entries will most likely have been used to train and test the model and/or "duplicated" entries in the test set inflated the metric because the model was predicting them twice correctly), but under the assumption of the data being representative, it is expected that in the real-world deployment the model will face similar circumstances.

5.3. Determining Next Steps

Considering everything done so far, our team decided it was appropriate to proceed to the deployment of the project. Although there are always additional steps that could be made -e.g., creating different black-box models to compare the results against the performance of our white-box approach or create separated models for the two different business goals - those would require more time and effort that in our understanding would be better spent testing the model applicability in real data. This decision was achieved after a thorough process that involved brainstorming of different possible directions, where none gained special traction as the overall opinion was that the model and process followed provided an elegant solution to the problem at hand.

6.Deployment

6.1. Deployment Plan

Firstly, it is not possible to have a good machine learning system in place if data does not have the proper quality. In this case, we seem not only to have problems with some variables (e.g., lack of timestamp) but we also felt that some more personal data could have been collected to improve the quality of any system that we may suggest being put in place.

Regarding the isolated predictive model created, it was constructed in such a way that it can be used since the moment the booking is reserved, as we attempted to not use variables that should not have been provided at that stage.

Concerning the system to implement, we believe that some principles of the MLOps would constitute good practice in this case. In this regard, this process should be seen as the first step towards a continuous organizational effort that should involve a continuous integration and deployment of the machine learning techniques, feed-forward to a constant monitorization process that should analyse the results and react according to triggers, that we will develop more in the next section.

Consequentially, we believe the next steps should involve the separation of the machine learning steps described in this report into different scripts (e.g., integration of features, data processing, among others) and its integration in a pipeline. This would allow for an optimization of the process in a way that the data could, in an automated way, be converted into useful business insight.

Additionally, and as part of the system we want to implement, our team also developed a prototype flask application, that can be later on improved as more budget is allocated to the project.

This application will enable Michael and his team to use the model created, having only to provide the characteristics of a booking and the intention in the classification. In this context, our application will allow Michael not only to figure out if he should contact a customer to offer incentives, according to our suggestion of contacting around 25% of the most likely customers to cancel, but also, to predict the outcome of a particular booking or even calculate the net demand if proving a CSV file with all bookings. In this sense, we believe that after every booking is reserved, it should go through our application, and Michael could decide to contact the customer to attempt to revert a highly probable cancelation, or it could just use the model to consider the number of possible cancellations there will be and overbook the hotel for that specific amount. As a suggestion, which would have to be economically validated, we would recommend that probably in low seasons it would be better to use the model to find the customer more likely to cancel and try to revert, while in high season it may be a better option to use it to overbook the hotel, as with time available (that our model provides) it would probably be easy to find a new customer that will pay full price.

6.2. Plan Monitoring and Maintenance

It is fundamental to monitor the various pipelines presented in the previous section, aiming to maintain or even improve the predictive ability of this particular machine learning model, so that it does not degrade over time. The first step towards this is to keep in mind the marketing efforts that may result from this and future projects, meaning that the fact that Hotel 2 will, for instance, try to prevent cancellations by contacting the most likely customers to cancel, should be incorporated in the process of net demand forecast. In this respect, we reinforce the proposed formula: “total bookings” – “predicted cancellation” + “predicted reconversions”, where “predicted reconversions” should be derived from a new predictive model constructed after enough data is collected.

In addition to the already mentioned, the use of the flask application should be carefully monitored to assure that the metrics used to evaluate the model keep providing acceptable results. In case the results fall below a certain level (which we should set a trigger accordingly), we should under the MLOps’s guidelines have in place a retraining process. This situation can be a result of different things, among which the change of the nature of the customers (with the parameters from new bookings differing substantially from the ones used to train the model) or even a conceptual drift, which would imply a change in the relation between the target and the predictive variables, that may be caused by exogenous shocks, such as the implementation of regulation over Online Travel Agencies or even the occurrence of a pandemic of unprecedented nature.

6.3. Conclusions and Brief Review of the Project

Although the review of the project should be discussed with the various stakeholders - in order not only to identify possible pitfalls or processes that might be improved, but also to keep track of the results of the deployment plan and if some major deviance occurred -, we believe that we can already present some insight in the form of general conclusions of the project.

Overall, we believe to have been able to provide to Hotel 2 a predictive model that respects the major business objectives purposed. In this regard, we provided the tools and a suggestion on how to “find” the most likely customers to cancel, while also providing a good forecasting model that goes in line with what Michael told us about both risks, of overbooking and having spare inventory, being very costly.

We also think that the planned deployment is well structured, meaning that we provided a framework that if followed will allow our model (and subsequent machine learning efforts) to continuously provide good results, keeping in mind that this process is not static in time, and should be checked regularly according to the recommendations provided for monitoring and maintenance. Lastly, we recommend Hotel 2 to purchase an upgraded version of the flask application, as this will allow more flexibility in its use and improve the general usability.

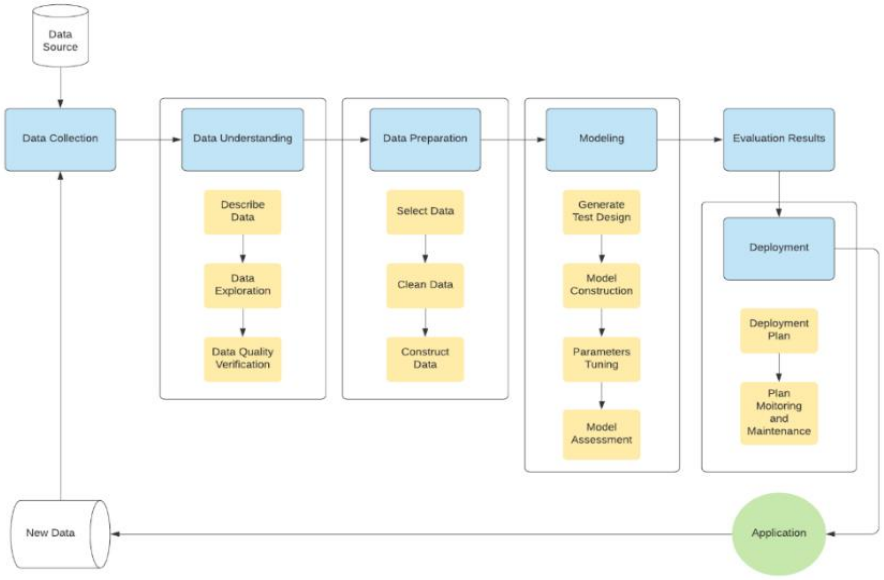

Figure 5: Brief overview of the process followed.

7. References

Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A. & Plöd, M. *MLOps Principles*. Retrieved from: https://ml-ops.org/content/mlops-principles

Brownlee, J. (2017, December 15). *A Gentle Introduction to Concept Drift in Machine Learning*. Retrieved from: https://machinelearningmastery.com/gentle-introduction-concept-drift-machine-learning/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0*. SPSS.