INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

CASE STUDY

Tasks for Course: DLMDSME01 – Model Engineering

CONTENT

1. TASKS

In this section, you can select one of the listed case studies to work on (see sections 1.1, 1.2, 1.3).

When working on your case study, please consider the task described in the respective case study itself.

In addition, each case study consists of a *.zip folder, which contains a *.csv dataset and *.txt files with further relevant information, like a description of the dataset. The *.zip folders are available in myCampus, under the section Case Study.

The course book and in-depth information from the internet or suitable reference books serve as a basis for the case study. For instance, data science community websites, like **kaggle** (https://www.kaggle.com/), are a good source of open databases and public notebooks. It is expected that additional relevant literature and current academic sources will be researched and used to develop the chosen topic. These sources must be correctly referenced in the bibliography; the correct referencing methods are described in the guidelines (cf. myCampus).

Case studies in the module “Model Engineering”, in particular, consist of conceptual and coding steps. Your observations and visualizations should be embedded into a final document; the format is described in detail in the guidelines (cf. myCampus). For the coding part, the programming languages **python** and **R**, which are the most popular in the data science community, are recommended. Moreover, the open-source web application **Jupyter Notebook** (https://jupyter.org/) is very helpful for an exploratory study as you can embed code, visualizations, and text into one single notebook file.

Do not forget to insert your well-documented code as an attachment to the final submission document since, via Turnitin, it is only possible to submit one document.

Note on copyright and plagiarism:

Please take note that IU Internationale Hochschule GmbH holds the copyright to the examination tasks. We expressly object to the publication of tasks on third-party platforms. In the event of a violation, IU Internationale Hochschule is entitled to injunctive relief. We would like to point out that every submitted written assignment is checked using a plagiarism software. We therefore suggest not to share solutions under any circumstances, as this may give rise to the suspicion of plagiarism.

1.1 Task 1: Credit Card Routing for Online Purchase via Predictive Modelling

Case description:

This is your first day as a data scientist at one of the world's largest retail companies. Already on your first day, you are invited to a meeting with important business stakeholders from the online payment department, who ask for your help: Over the past year they have encountered a high failure rate of online credit card payments. The company loses a lot of money due to failed transactions and customers become increasingly unsatisfied with the online shop.

Such online credit card payments are performed via so-called payment service providers, referred to as "PSPs" by the business stakeholders. Your company has contracts with four different PSPs and pays transaction fees for every single payment.

The current routing logic is manual and rule-based. Business decision-makers, however, hope that with predictive modelling and with your help, a smarter way of routing a PSP to a transaction is possible.

Project Aim:

Help the business to automate the credit card routing via a predictive model. Such a model should **increase the payment success rate** by finding the best possible PSP for each transaction and at the same time **keep the transaction fees low.**

Data Set:

The data set and all relevant information from the business side (name of PSPs, transaction fees) are given in a separate *.zip folder, which is available in myCampus, under the section Case Study.

Task Description:

The task consists of both coding and conceptual steps. Here is a list of tasks, which should be included in your final document:

- **Structure the project** via the CRISP-DM or Team DS methodologies and give a recommendation of how a git repository for the project could look like. Note that you do not have to structure your final code according to your git-repository proposal.
- **Assess the quality** of the provided data set. **Prepare and visualize** your findings of the initial data analysis in order that business stakeholders can understand them in a clear and easy way.
- **Provide a baseline model** as well as an **accurate predictive model, which fulfills business requirements**, i.e. increase credit card success rate and keep fees low.
- In order that the business places confidence in your model, discuss the importance of the individual features and make the results of the model interpretable. Moreover, a sophisticated error analysis is very important for the business to understand the drawbacks of your approach.
- In the last step of the project, give a proposal of how your model could be used by the business in everyday work, for instance, via a graphical user interface (GUI).
- Finally, do not forget to attach the code to the final submission document.

1.2 Task 2: Automation of Standby Duty Planning for Rescue Drivers via a Forecasting Model

Case description:

As a data science consultant, you are invited to work on a project within Berlin's red-cross rescue service. The HR planning struggles with the current standby-duty plan. They ask for your expertise on predictive models in order to improve the current planning logic.

Every day a certain number of rescue drivers are on duty. However, due to short-term sickness of rescue drivers or an unusual amount of emergency calls often more drivers are needed than initially expected. Hence, a certain number of standby-drivers are kept "on hold" and activated when needed. Therefore, in the current approach 90 rescue drivers are daily kept on standby.

As colleagues from the planning claim, there are seasonal patterns - for instance, during winter months more rescue drivers call sick – which are not incorporated into the current approach. Moreover, sometimes there are not enough rescue drivers even when all 90 standby-drivers are activated so that drivers are called for work even on their days off.

It is important to know, that the duty plan must be finished on the 15th of the current month for the upcoming month. This means, for instance, that the duty plan for November must be finished on the 15th of October.

Project Aim:

Help the HR department with planning to estimate the **amount of daily standby rescue drivers via a prediction model** more efficiently. Here, *efficient* means that the percentage of standbys being activated is higher than in the current approach of keeping 90 drivers on hold. It also means that situations with not enough standbys should occur less often than in the current approach. Note that the plan must be finished on the 15th of the current month for the upcoming month.

Data Set:

The data set and all relevant information from the business side (sickness dataset, number of emergency calls per day) are given in a separate *.zip folder, which is available in myCampus, under the section Case Study.

Task Description:

The task consists of both coding and conceptual steps. Here is a list of tasks, which should be included in your final document:

- **Structure the project** via the CRISP-DM or Team DS methodologies and give a recommendation of how a git repository for the project could look like. Note that you do not have to structure your final code according to your git-repository proposal.
- **Assess the quality** of the provided data set. **Prepare and visualize** your findings of the initial data analysis in order that business stakeholders can understand them in a clear and easy way.
- **Provide a baseline model** as well as an accurate **predictive model, which fulfils business requirements**, i.e. the amount of activated standby-drivers is maximized, but having not enough standbys is minimized.
- In order that the business places confidence in your model, discuss the importance of the individual features and make the results of the model interpretable. Moreover, a sophisticated error analysis is very important for the business to understand the cases in which your approach could potentially fail.
- In the last step of the project, give a proposal of how your model could be used by the business in everyday work, for instance, via a graphical user interface (GUI).
- Finally, do not forget to attach the code to the final submission document.

1.3 Task 3: Efficient Flight Operations: East Carmen Airlines via Machine Learning Flight Prediction Model

Case description:

Welcome to East Carmen Airlines! As a strategic data scientist, you are confronted with many use cases to digitalize the airline business. Most important for airlines are accurate predictions of flight arrivals during a day of operations. When it comes to passenger handling, such predictions would admit a proactive steering, in order that the passengers reach the connection flights and therefore save a lot of money.

Your model should be applicable to each aircraft in East Carmen's fleet. Let us take as an example the aircraft with the registration name EC-LPD. At the scheduled departure of the earliest morning flight from East Carmen's hub Madrid the model should predict all flights of EC-LPD on the given day. Many business stakeholders do not trust the robustness of such a model. They claim that for a given aircraft, one could predict the first flight of the day with a decent error, but the errors of second and third flights will be huge. It is now your task to convince the business stakeholders that with a well-prepared data set and meaningful features you can deliver robust flight predictions.

Project Aim:

From different departments, business analysts have collected a lot of information on possible features for flight predictions. Now it is your task to use these data to deliver a **prototype forecasting model for all flights of a given aircraft** at the scheduled departure of the first flight of the day. For instance, for an aircraft with registration name EC-LPD, first flight of the day (FFOD) scheduled at 5:30 am from MAD to VIE and with the connection chain MAD-VIE-MAD-CDG-MAD-LHR-MAD your model should predict the successive landing times in VIE, MAD, CDG, MAD, LHR, MAD. The colleagues from daily operations are especially interested in how the **confidence level of such a model evolves during a day of operations.**

Data Set:

The data set and all relevant information from the business side (operational-, weather data per flight, fleet list) are given in a separate *.zip folder, which are available in myCampus, under the section Case Study.

Task Description:

The task consists of both coding and conceptual steps. Here is a list of tasks, which should be included in your final document:

- **Structure the project** via the CRISP-DM or Team DS methodologies and give a recommendation of how a git repository for the project could look like. Note that you do not have to structure your final code according to your git-repository proposal.
- **Assess the quality** of the provided data set. **Prepare and visualize** your findings of the initial data analysis in order that business stakeholders can understand them in a clear and easy way.
- **Provide a baseline model** as well as a **prototype predictive model, which fulfills business requirements** described above in the "project aim".
- In order that the business places confidence in your model, discuss the importance of the individual features and make the results of the model interpretable. Note that it is especially important for the business colleagues to understand the error of predictions for successive flights in the chain.
- In the last step of the project give a proposal of how your model could be used in everyday work, for instance, via a graphical user interface (GUI). Discuss possible relevant information, which is not yet included into the data set and which could help to improve the forecast.
- Finally, do not forget to attach the code to the final submission document.

2. ADDITIONAL INFORMATION FOR THE EVALUATION OF THE CASE STUDY

When conceptualizing and writing the case study, the evaluation criteria and explanations given in the writing guidelines should be considered.

Identification: How is the business problem translated into a data science framework, i.e. which data science use case can be established from the given business problem?

- Example: "For the problem at hand we first apply a supervised learning classification algorithm. From the predicted probabilities we find a rule-based logic for the system."

Concepts: Application of relevant concepts learned throughout the course.

- Examples are:
  - how to perform the initial data analysis, like calculating correlations
  - which models are used for predictions (Linear Regression, Random Forest…)
  - which techniques are used to select import features
  - how to prevent overfitting (train hyperparameters, use regularization…)
  - …

Analysis: Deliver a sophisticated, detailed analysis of the data and the modelling approach.

- Example: "In our initial data analysis we observe that for Argentina the number of customers decreased by 50%, whereas in all other countries the decrease is on average 10%. This means that in Argentina we face a huge problem with our business."
- Example: „We observe that choosing a random forest classifier increases the overall accuracy of the model by more than 5%. Moreover, when we apply a trigonometric function on the temporal features the accuracy increases by another 5%."

Conclusion: Describe the important results of your analysis in a clear and quantitative way. This goes hand in hand with the analysis. How can your model be used in everyday work? Are there any improvements necessary until the model goes into production?

- Example: "From the analysis of the model we conclude that our new data-driven approach outperforms the old, rule-based approach in our test set by 20%. We did not find any situation where the rule-based approach gives better results. Therefore, we are confident that the prediction model will lead to a significant cost reduction. Our conservative estimate for the next two months is a cost reduction of more than 6% due to the prediction model."

Formalities: Adherence to the guidelines.

Accuracy: Correctness of spelling and punctuation.

Language: Quality of the linguistic expression and adequacy of language style for scientific work.

3. TUTORIAL SUPPORT

Students have the option to make use of any one of several opportunities to get support for their case study analysis with the course tutor. Taking advantage of these opportunities is the responsibility of the student and the use of these services is voluntary. It is possible to contact the tutor regarding formal and general questions about working on the case study. Please note: a review of outlines and aspects of the presentation is not intended here, since the student's ability to work independently is part of the evaluation and counts as a part of the overall assessment. There are however general tips for developing the case study to help you getting started.