Report E1 - Homework 10

Project Title: ***Improving the quality of museums data***

Group Members: Till Wenke, Florian Nebenführ, Daria Eckert

https://github.com/TillWenke/museum_item_classification/ - is already public

Business Understanding

Identifying Your Business Goals:

Background

Our project is part of the AI government initiative in Estonia, as part of this the National Heritage board that oversees the Estonian museums is our client These museums have a vast collection of items that have been previously examined and described using several different descriptive factors from location to originality. These artefacts have however not all been formally labelled into categories such as photograph, painting and coin etc. In order for these items to be effectively used and presented within the Estonian museums information system MuIS (the collaborative online collection system), the gaps in the database of unlabeled items need to be filled. Doing that would in turn help the museums provide a meaningful and accessible educational platform.

Business Goals

The National Heritage Board wants to “improve Estonian museums information system and preserve Estonia’s cultural heritage”. More well prepared, accessible data about museum items could be of interest for different stakeholders such as citizens, museum workers or scholars. To achieve this goal, we are asked to develop a machine-learning model that can predict the categorical label of an artefact based on these descriptive factors for the remaining unlabeled items in the museum dataset and thus speed up their otherwise manual categorization process. Although this data with missing labels is not presented to us.

Business Success Criteria

No defined - we can only derive more specific data-science/ ML-success criteria from the business goals. Labelling the items wrong in a public catalogue (or at least it should occur in very rare occasions, lets say 99 %). Another goal could be to add a certainty to the predictions and keep the number of uncertain items to a manually manageable amount. As a consequence the proportion of certain (most likely correct labelled) items can be lower that 99%. For now and for our project, no concrete criterium is defined by the “customer” although the models’ predictions on kaggle are evaluated using accuracy scores and the currently best result is around 0.912 from a previous competition. As there is no further benchmark we could define our success as being noticeably better (let us say 0.92-0.93) than this previous approach. There are no insights provided about the type of model used to reach this score consequently we cannot estimate if our approach is completely different and thus promises results that are not just slightly better (let us say above 0.95).

Assessing Your Situation:

Inventory of Resources

The following resources are available to us for our project:

- The public dataset from the Kaggle competition https://www.kaggle.com/competitions/caps-in-museums-data-part-2/
- The google translation tool - to help us gain an understanding for the contents on the dataset as the contents are in Estonian and we need to understand them in the preparation phase and maybe run more sophisticated NLP on them.
- Tool for tracking the models while hyperparameter search if necessary - https://wandb.ai/
- Kristjan Eljand the competition owner - ready to answer our question about the dataset.

Requirements, assumptions, and constraints

- **Requirements**:
  - Kaggle competition completed by the 7th of Dec and project by 12th Dec.
  - Submitting the predicted classes only for all test-instances.

- **Constraints**:
  - Dataset written in Estonian - need for translation for understanding but modelling maybe based on the original Estonian dataset.
  - Really sparse dataset, some features barely contain any values.
  - Although presented in tabular format there are a lot of features containing longer paragraphs of text which in our opinion could contain the most valuable information (e.g. sometimes the label/ type is clearly stated in them). Consequently, we would have to use NLP besides classic ML approaches.
  - Some targets are very rare so it will be difficult to have examples to learn from, we could think about not aiming to label them correctly.

- **Assumptions**:
  - Once translated it would be most convenient to work further with the translated dataset, but we cannot assess how much or if information got lost during the process. So firstly, we would assume that the English dataset will help us in using existing models trained in this language but we would also try working with the original dataset and compare the result, especially when we see that we fall short.

Risks and Contingencies

- As one can assume from the existence of unclassified items in Estonian museums, this catalogue was not designed with data science and especially clearly distinguished features in mind. This means we had to do a lot of work in feature engineering, casting the present information into more useful features.
- With 38 features most of which are categorical (and many categories) the dataset is getting quite wide quickly when one-hot-encoding the features. Besides that many of the categories are distributed exponentially so it could be that most of them are not really relevant to draw general conclusions. We would have to balance between not including some irrelevant categories (most likely for runtime and model performance reasons - not yet sure if we will encounter them) and losing potentially relevant information.

Terminology

We've decided to leave it out as it isn't as relevant to us.

Costs and Benefits

As of now we are working for the good of the Estonian society. If we feel like sophisticated language models can help then there would also be “real” costs related to making our predictions, which should be neglectable when making predictions just once on a (the real unlabelled dataset from the Estonian museums) dataset of this size.

Defining Your Data-mining Goals:

Not relevant other goals already defined above in Business Success Criteria and Goals.

Data Understanding

Gathering Data:

Outline Data Requirements

To build a model that can predict museum item categorical labels based on descriptive factors we need a training dataset which has the same descriptive factors for each item and their respective labels. This has been provided by the Kaggle dataset.
In order to build models using this dataset we need to prep the data to be more suitable through hot encoding, mapping and potentially keyword extraction.

Verify Data Availability

For our project we are choosing to stay within the scope of the provided kaggle data for the museum dataset. Although one could think of extending it with more items from https://www.muis.ee/en_GB/ if needed.

Define selection criteria

Not applicable

Describing Data & Exploring Data:

Our training dataset consists of 14000 instances and has 36 text based original descriptive features. While exploring our data we noticed several things:

- The dataset is written in Estonian (except the feature names) but also contains other languages that influenced Estonia throughout history (Russian (also Cyrillc), German). We'll need to pay attention to this especially during translation.
- Interestingly we can see different kinds of features some obviously seem very helpful even to humans when guessing the type of an object such as the technique used or colour whereas others seem quite irrelevant such as a "participant" related to items or before_Christ which only contains no and NaN values.
- The dataset has a lot of missing values - a lot of columns are mostly NaNs eg. commentary, parish.
- Largely text based values for each feature - for the longer texts it is hard to turn them into categorical data eg. commentary, additional text.
- There are a lot of possible labels for items in the dataset - some only have very few cases (nukk, tunnistus) making them hard to predict while others appear very frequent (some sort of photo in about 6000 instances). So we are dealing with a highly imbalanced dataset and will need to come up with some coping-mechanisms.

- Some feature-values are not provided in a uniform format such as "start" (years/ dates), others contain multiple different information (eg country_and_unit should better be split up) or not only contain categories of values but also lists (eg. material, technique).

- There is one interesting feature: full_nr which contains a museum specific identifier for the items which is clearly not useful as it is unique but it is already split up into its parts which seem to be partially inconsistent (sometimes numbers or characters used dependent on the museum). We assume that it would make sense to have some categories/ sections of museums hidden in this identifier that we cannot see yet but it is very likely that ML models can draw those relationships.

- There are barely any numeric features despite values for different measurement about the items. Interestingly only one measurement is given for each (eg. either weight or size). Those values their units and meaning is spread over 3 features so it is our task to establish more meaningful numeric features such as “length in m” so that the model can draw conclusions without sense-making from the previously unlinked features.

- Some feature values are also compromised and demand cleaning such as extra white spaces or “<>” such as country_and_unit or event_type.

Verifying Data Quality

The data provided has the same structure and descriptive features as the data we would like to later predict, although some feature-categories may only appear in train or test set which should be kept in mind.
The issues mentioned above will need to be addressed via the following:

- Converting as much of the features as possible to hot encoded categorical data, for some features this could mean taking the top x examples of that feature and hot encoding these to limit the overall number of features to the most relevant.
- Splitting up some features to more useable sub features.
- Especially we will divide the feature into those that contain categories or numbers and those that only contain extensive text and treat them differently which means classical ML for the first and NLP for the latter.
- Renaming certain values
- Grouping similar values together for more useable sizes

Planning Our Project

Tasks	Time spent my team members	Methods/Tools
Stage 1: Data Understanding • Understanding of Column names and variables - including translation from estonian to English • Counting NaNs • Identify for each column whether it needs to split, mapped, hot encoded or adjusted	Till: 5 h Daria: 2 h Florian: 1 h	-Jupyter notebook/python for reading and manipulating data - pyhton library relying on GoogleTranslate
Stage 2: Data Preparation		-Jupyter notebook/python for reading and manipulating data
• Combine train and test dataset	Daria: ½ h	
• Cleaning all columns of punctuation/formatting errors eg <>	Florian: 5h	
• Learn to splitting certain columns into smaller more usable features	Till: 9 h	
• Finish and clean the translation	Till: 1 h	- pyhton library relying on GoogleTranslate
• Mapping some columns to 0/1 and hot encoding the categorical data columns	Daria: 5h	
Stage 3: Training Our Model (Not yet fully decided)	Tasks haven't been divided yet Till: 18 h Florian: 12h Daria: 6 h	
• Adjust dataframe for different models		-Jupyter notebook/python for reading and manipulating data
• Split into train/val for cross validation		-Jupyter notebook/python for reading and manipulating data

• Decide on models to try - potentially including deep learning		-The python library of models used in the lecture
• Dig into NLP for text-based features and combine predictions/ extracted keywords with predictions from (most likely random forest classifiers)		- NLP: gpt3 or open/ unlimited alternative or any lighter approaches/ models such as https://ner.tartunlp.ai/api (in case we need to switch back to Estonian)
• Checkout usually well-performing sophisticated models such as xgboost or Tabnet and looking into methods to combine models.		
• Train (cross val) on different models		-Jupyter notebook/python for reading and manipulating data -https://wandb.ai/site for tracking model parameters during tuning -The python library of models used in the lecture
• Model/ state certainty for predictions		-https://wandb.ai/site for tracking model parameters during tuning
• Further tune best model		-Jupyter notebook/python for reading and manipulating data -https://wandb.ai/site for tracking model parameters during tuning
Stage 4: Presentation/ Report (Not yet fully decided)	Tasks haven't been divided yet Till: 6 h Florian: 12h Daria: 18 h	
• Run our best model on the reserved test set		-Jupyter notebook/python for reading and manipulating data
• Analyse our models success for different labels and features		-Jupyter notebook/python for reading and manipulating data
• If applicable and necessary/ possibly investigate certainty and what missing feature values could help to make a		

prediction		
• Complete our poster on the conclusions of our project		-Jupyter notebook/python for reading and manipulating data - relying on python visualisation libraries if necessary