Practical Machine Learning & Deep Learning Project Proposal

Team members: Kamil Sabbagh, Fadi Younes, Rafik Hachana

Innopolis University, Fall semester 2022

Topic: *Apartment condition classification from images*

Goal and motivation of the Project:

Motivation:

Since the rise of the World Wide Web, the idea of markets and trading drastically changed to the point that trade nowadays does not require human interaction between buying and selling parties. In fact, most trade contracts can now be done online with no middle party involved. However, every change comes with its downsides. The most straightforward downside of this is the significant increase of scammers with no way to reveal them. And the repercussions of this issue drastically increase for expensive trade deals, such as real estate deals. Most scammers tend to overestimate the apartment condition thus, the client may waste time to find the perfect apartment, or inevitably settle for a less convenient apartment.

Unfortunately, the manual filtering process for such an issue consumes plenty of time and money. But automating the filtration will help increase the clients' satisfaction and fight the deliberate attempts of scamming. The goal of our project is to determine how Deep Learning models can be an effective tool in the automation of this process.

Modeling:

As previously noted, manual filtering would consume lots of resources and therefore is not a practical solution. So, in order to find the answer, we aim to build a deep learning model that classifies apartments based on their condition into valid or scam (overpriced) apartments judging by their images and information. We aim to provide a classification model with a minimal working state that can be later improved (by tuning and further training) to demonstrate the feasibility of our idea.

The objective of the classification model is to identify apartments with overestimated condition among the offered apartments, the input for the model would be both non-tabular data (such as images) and tabular data (such as price, area, etc…). The output of the model would be 'Overestimated' for overestimated apartments, and 'Normal' for accurately-estimated apartments. Ideally, the model will have the flexibility to deal with data loss (in case some apartment offers do not provide all the data usually used for input).

Data collection:

Data will be collected from real estate websites (Otodom & Morizon), which are mostly active in the Polish real estate market. The data will be collected by scraping the target websites to collect the data samples. After the cleaning and preprocessing stages, further observations will be extracted from images and added separately to our datasets.

Desired output:

The desired output of this project is an insight on the efficiency of Machine Learning models in identifying scam operations in real estate web markets, as well as a basic working software that can be used by the end-user to perform the apartment classification. In particular, we are only interested in providing a proof of concept, meaning that it is appropriate to conclude that AI can be an effective tool for this purpose if we can build a basic solution with an accuracy higher than 60%. The deliverables at the end of this project are: The models used in the comparison, the best performing model, the code used for scraping, preparing and cleaning data, final data that was used to train and validate the models, as well as all the required documentation (4 progress reports, 1 final technical report, and a final presentation on the whole project).

Data Description:

The Dataset to be obtained is planned to contain 11 observations: Source website of each element, year of build, price of the apartment (in Polish złoty), area, price per square meter (in Polish złoty), number of rooms, address, link on the source website, description, renovation condition, and the main image of the apartment in the ad page. The data was collected by scrapping the websites Otodom & Morizon which are famous for being a great real estate market in Poland. A MongoDB database is used to store the data, and there will be two collections, one for each website.

It is also necessary to point out that some appartments offered on these websites are not yet built (they are usually built by contractors after the deal is done and the payment is received, and the final handover is after a few years). Such offers usually provide the architectural plan for the apartments to be built in the attached images. We will exclude these apartments from our dataset as they can't be predicted to be in a good or bad condition before they are built.

Software:

The software is to be developed solely in python. The following libraries will be used for several purposes:

1. Pymongo, Pandas, and Numpy will be used for Data collection, cleaning, and manipulation.
2. Sci-kit Learn, and PyTorch will be used for Model Building. (We might also include other libraries like Tensorflow, HuggingFace depending on the availability of pre-trained models).
3. OpenCV will be used for Image Processing.

Analysis Plan:

The plan for this project consists of three stages:

The first stage is Data Collection. Data must be scraped from the source websites and stored in a MongoDB database. This process includes both writing and implementing the scraping code (until enough data is stored) and is planned to be done within two weeks.

The next stage is Data Cleaning & Preprocessing. Some records in our dataset may be duplicated, inaccurate, or incomplete (missing data imputation), and will therefore require fixing. In the second part of this stage, we will also fix any structural errors such as misspellings, Incongruent naming conventions. Both parts of this stage are planned to be done in two weeks.

The third and final stage is model building. This stage is also expected to take two weeks, in which we will research all the Deep Learning models that can solve this task efficiently, and train the best models on the dataset acquired. Then pick one model or combine the results from multiple models to make the final solution.

The last week will be left for any issues that may come up during any of the three earlier stages that weren't considered earlier.