UDACITY

PROJECT

Capstone Proposal

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

Hi,
Good job so far! This is a challenging problem as off-the-shelf classifiers are not going to produce reasonable results due to high imbalance, so you need to resort to anomaly detection and other techniques. In general, you have a solid proposal, there are just some aspects to be further discussed. I've added some suggestions related to alternative ways to cope with the imbalance, such as cost-sensitive methods (Gaussian NB) and oversampling, however these may assume a supervised learning approach, which is likely not your preferred option at the moment. Anyway, you may consider these for your benchmark model.

Best Regards.

Project Proposal

Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.

Awesome

- Really good introduction to fraudulent transactions! You have added references to backup your arguments and also introduced academic research for this problem.
- Good to know about your motivation and background.

Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant potential solution. Additionally, the problem is quantifiable, measurable, and replicable.

Awesome

- Important characteristics of the machine learning problem were clearly mentioned, e.g., the problem is highly imbalanced, thus you will need to resort to anomaly detection.

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

Awesome

- The origin of the dataset (Kaggle competition) was presented and a reference to it provided.
- The total number of transactions as well as the distribution between fraud and non-fraud was presented.
- You discuss details about the input features, such as how they were anonymized and how many there are.

Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.

Suggestion

- This seems like a reasonable approach, still, since the input was already reduced through PCA it is likely that the features are already quite representative. So, if you end up with feature importances obtained through a Decision Tree (or ensemble versions such as Random Forest, which are more stable by the way) that are really similar this might be the

reason for that. Also, notice that there is a high imbalance in the dataset and you may need to verify the generated tree just to be sure that it didn't output a model such as "always predict legit", which would definitely give you not useful feature importances.

Required

- What learning methods will be your main solution to the problem? Why you have selected them? In the Project Design section you mention EM, but it is not clear why you are going for an unsupervised learning technique, you should explain that in here. Given your background you might be considering using an unsupervised learning technique as it can be deployed in a dataset without labeled instances easily, if that is your reasoning, please explain that in here.

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

Suggestion

- Why are you going to create an initial benchmark model based on the approach used in the "Titanic Survival Exploration"?
- I believe you meant "logistic regression" instead of "linear regression" in here.
- If you only apply these supervised learning techniques the result will be the same to all of them, i.e., "always legit". Therefore, if you want a reasonable benchmark focus in methods that are able to be able to predict the minority class fraud. For example, you may try Gaussian Naive Bayes and tune its `class_prior_` hyper-parameter, which controls the "weight" given for each class and may be used to better model this highly imbalanced problem (methods that allow you to weight classes are known as "cost-sensitive" approaches).

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

Awesome

- AOPRC is a reasonable metric for this problem and you explain why that is the case well.

Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

Suggestion

- This is a form of undersampling:

  "GENERATING A SET WITH ALL THE FRAUDULENT TRANSACTIONS AND THEN ADDING LEGIT TRANSACTIONS ABOUT 4 TIMES THE FRAUDULENT TRANSACTIONS."

  In practical terms, you will throw away part of the legit transactions to better balance the data. There are other undersampling and even oversampling (i.e. create synthetic fraud transactions) techniques that you might be interested on for this problem in this python library.

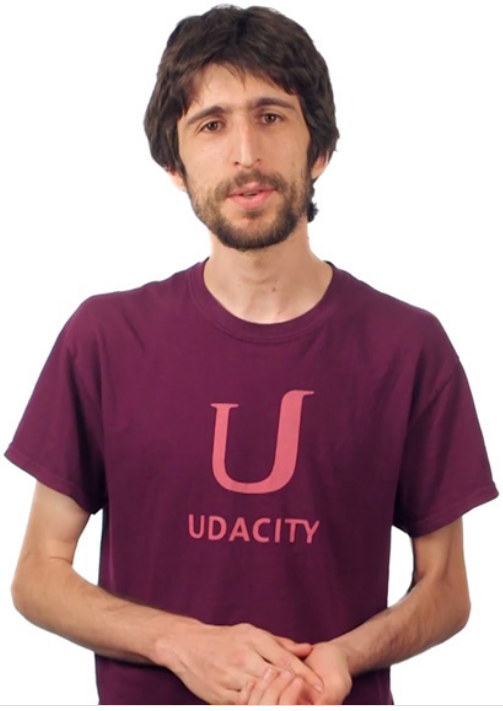
Proposal follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used and referenced are properly cited.

Awesome

- The proposal is well structured.
- Resources were properly referenced and cited.

RESUBMIT

DOWNLOAD PROJECT



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH

Rate this review

Student FAQ Reviewer Agreement