Machine Learning Project for InvesTarget

Description of Deal Evaluation Machine Learning Model

January 15th, 2017

InvesTarget is a leading cross-border investment bank co-founded by professionals from global top-tier investment banks, consulting firms and investment funds.

InvesTarget has a deal platform currently with more than 1,200 active projects.

通过 Machine Learning（以后简称 ML）主要是提供用户两个方向的主要需求
We want to achieve two main objectives utilizing Machine Learning (hereafter ML).

项目评价以及相关资料爬取

Project Evaluation and relevant information collection using web crawling

As we have a considerable deal flow volume from all over the world, we hope your team can help us develop a ML algorithm to evaluate the projects. For each project, we usually have a 20-page deck and/or a one/two-pager executive summary that contain basic information about the target.

- Dimension 1

在为平台每一个用户做项目推荐的时候，通过用户的个性化标签对用户进行项目推荐的"千人千面"。
一个维度是比如我们平台上有 5000 个项目和 4000 个用户，我们要求用户为项目打分。每个项目分属于不同行业领域，我们可以看出 A 用户和 B 用户似乎更倾向与医疗行业项目，而 C 用户和 D 用户似乎更倾向新农业项目。并且没有一个用户给所有的项目都打过分。我们希望构建一个算法来预测他们每个人可能会给他们没看过的项目打多少分。也就是 AI 现在的一个基本研发方向建立人工智能的推荐系统

When recommending the projects to each user of the deal platform, we hope to utilize user's expertise, based on their tags such as Healthcare, Robotics and Manufacturing, as they are all experienced institutional investors specialized in different sectors.

For example, there are 5,000 projects and 4,000 users on the platform and we ask each user to give each project a score when read the project profile. Every project is in a certain sub-sector. We can see User A and B are more interested in Healthcare projects and User C and D are more interested in Modern Agriculture. None of the users would grade all the projects. Therefore, we hope to develop an algorithm to predict that how would each user score a project that they didn't read.

- Dimension 2

另外一个维度，我们希望 AI 能够找出某个领域（比如找寻欧洲汽车变速箱领域比较有竞争力的企业），通过收集这些公司公布在互联网上所有和该公司相关的结构化和非结构化数据，最终来生成该行业的好公司列表。同样的这里可能大量牵涉到自然语言分析（NLP）以及大量的数据清洗工作，同时我们会提供大量最终成交的好项目来为 AI 作为宝贵的训练数据，并且还需要让 AI 判断这个公司的项目最终愿意接受股权融资或者并购的可能性（这个需求更多是针对海外并购项目的投前发现）

We are looking to train the learning model via supervised learning.

When we are sourcing deals in specific sectors, for example leading gearbox manufacturer in Europe, we hope to utilize both structured, e.g. financials, and unstructured data, e.g. business overview, management profile, operational plan, investor background, even video like interviews published on the internet, of the target company both from the materials and the data collected from web crawling. This might heavily involve Natural Language Processing and data cleansing. We will provide numerous examples of successfully closed deals to train the model. The set of training examples will be previous cases of investments, such as private placements, mergers and acquisitions. Eventually the model should be able to evaluate our new projects with a probability of success rate of being invested and acquired.

通过这个两个维度的结合最终会给项目打分，保证项目分数的客观性，同时将分数最好的项目推荐给对该行业感兴趣的用户。简单来说就是我们平台上的项目都不是分析师推荐的而是机器推荐的。

Combining the two dimensions, we assure the objectivity of the project score and recommend the best projects to the users based on scores. We are looking to replace our analysts' role in recommending projects with the ML model.

最终模型算法的使用方案

The application of the final ML model

我们希望可以通过 AWS 的云服务，将贵团队的机器学习算法和技术通过 API 接口的形式提供给国内的应用程序，让其具有预测能力。

We hope to utilize Amazon Web Services to apply the machine learning algorithm to our app and website via API.

1. General Description

The project aims to model investment projects and evaluate them by an objective score. To achieve this goal, both Computational Intelligence and Computer Engineering technologies can be used. The design and implementation of Computational Intelligence algorithms could be the core of the project. Meanwhile, Computer Engineering technologies could also be crucial because we should use them to obtain sufficient data and provide a satisfying interface to users.

In this section, the general idea of these two aspects will be analyzed respectively.

From the perspective of Computational Intelligence, we should use machine learning methods to design the model. Machine learning is probably the most popular method currently used by researchers and corporations. In the field, the two requirements of the project might be regarded as customer modeling (requirement 1) and supervised classification respectively. Although they are popular problems and existing methods have been designed towards them, given the particular condition of the project, the machine learning algorithm designing of this project is still non-trivial. For the requirement 1, the problem is that the training data is sparse. Current state-of-the-art customer modeling methods often utilize Neural Networks/ Deep Learning (such as Google and Facebook). However, fitting a Deep Learning model need millions if not billions of data, which is not applicable in our case. Consequently, a method which could specifically solve this problem should be designed. For the requirement 2, the problem is the relationship between features and the target output (willing to be merged) is unknown. Thus, to solve this problem, feature analysis and model selection are needed, which could take a relatively long time. Several corresponding potential solutions are provided in following sections.

From the perspective of Computer Engineering, there are two important tasks: web crawler and integrated program designing. With web crawler, it will be possible to collect data from websites. This procedure is crucial for the requirement 2, although collecting and cleaning the data could also be time-consuming. Another important topic is to design an integrated program, which could easily be used by any user even does not have any knowledge in computer. While these topics are important for our project, existing methods could be suitable and it is relatively easy to achieve the functions.

2. Methodology

1) Requirement 1: Customer Modeling

As stated before, the requirement 1 could be considered as customer modeling with sparse data. The difficulties in this problem mainly lay in following aspects:

a) The Machine Learning model is likely to overfit.

With only a limited number of data (which implies that only a limited amount of information), the model is likely to become overfitting meaning that the model which performs well on training data might have poor performance on test data.

b) Outlier Problem

An outlier is an observation point that is distant from other observations. In machine learning problems, we often treat outlier data as wrong or not representative data. Generally, if the data amount is considerable, the influence of outlier is not significant. However, in our case, because our data is sparse, this problem becomes serious.

c) Noise Problem

Noise is the colloquialism for recognized amounts of unexplained variation in a sample. Data with noise will be variant from what it should be but not too far (contrast to Outlier). Noise could be caused by various reasons: calculation, storage of decimal places and unpredictable environment influence. With a small amount of data, noise could affect the model with a higher level.

2) Requirement 2: Supervised Classification

Comparing to the requirement 1, the requirement 2 is mathematically more clear and could be achieved by existed approaches. However, practically, this topic is more time-consuming because this is not a very 'elegant' problem.

Mathematically, the main obstacle of the problem is that the relationship between the input and the output label is ambiguous. Although machine learning could capture arbitrarily complex input-output relationship, a non-explainable model is precarious for commercial project because it could not handle neither any potential changes in situations nor any failure reported by users. Thus, probabilistic model is highly recommended for this problem. Using Bayesian method, the output of classifier could be represented as a posterior probability, which could imply the sense of "uncertainty" for the decision. For example, if the classifier posterior probability for the willing of been merged of a company is 0.6, then our result is in low confidence and we probably should give this project a lower mark. In contrast, if the posterior probability is 0.9, then it should be a nice fit for the investment. One popular probabilistic classification method is the Gaussian Process Classification, which is also the method mentioned for the requirement 2.

Another alternative machine learning method could be Reinforcement Learning. Reinforcement learning is closely related to the Bayesian Decision Theory. In the reinforcement learning, each decision will be associated with a Utility which means the potential gain/loss of the decision. Reinforcement learning could find the optimal decision given the condition and because the Utility could be regarded as potential income/risk in our project, it could be an ideal fit for the project.

As it is stated before, the main difficulties of this problem actually lay in how to obtain data and analyze features. Comparing to the 'exquisite' mathematical methods in algorithm, there are no elegant approaches in this problem and we have to search and analyze manually. Web crawler could help us to obtain the data, but because of the large amount of webs we have to go through, it is still time-consuming. The collected data might be in unsuitable format and the information could be in verbal expression, thus data normalization and natural language processing (NLP) are required. Finally, we need to analyze whether the data is appropriate for training the model. Consequently, the requirement 2 may take a longer time in the plan.