A machine learning model for startup selection

Khadjimurad Gamzatov gamzatov.kha@phystech.edu
Daniil Merkulov daniil.merkulov@skoltech.ru

Project Proposal

Используя набор данных о стартапах из Crunchbase и United States Patent and Trademark Office (USPTO), в этом проекте разрабатывается модель машинного обучения для прогнозирования успешности стартапов, то есть поднимутся ли они до Series B и выше или потерпят неудачу. При использовании большого набора функций определяется accuracy прогнозов результатов запуска. Этот проект предполагает, что венчурные компании могут извлечь выгоду из использования машинного обучения для отбора потенциальных инвестиций с использованием доступной информации.

1 Идея

Согласно статье [1], автор которой ссылается на Gatner и говорят, что что к 2025 году 75% венчурных фондов будут использовать AI для принятия решений об инвестициях. Так же есть Гарвадское исследование [2], в котором сравнивали построенную ml модель с 200 венчурными ангелами. ML модель работала прибыльнее, чем начинающие или средне-опытные инвесторы, но хуже, чем очень опытные. Но все равно полезно, так как модель находит инсайды, и гибридный подход (предсказания от модели + человеческие решения) - самый выгодный. Разработаем свою ML модель на общедоступных данных, рассмотрим различные алгоритмы и значения метрик, которые они выдают и выберем лучшую.

1.1 Problem

Задачей проекта является разработка модели машинного обучения для прогнозирования успешности стартапов, а именно поднимутся ли они до Series B и выше или потерпят неудачу.

1.2 Data

Данные, необходимые для проекта, будут получены из Crunchbase и United States Patent and Trademark Office (USPTO). Большинство признаков, которые модели используют в качестве входных данных, будут получены из Crunchbase.

1.3 Method

К нашим признакам применим train-test-split (80/20), а затем тестируем различные модели, включая XGBoost, Logistic Regression, KNN, SVM и Random Forest.

2 Outcomes

В качестве результата проекта будут обученные модели машинного обучения, метрика accuracy для каждой из них.

3 Литературный обзор

В статье [3] использовались довольно простые модели, но с несколькими интересными мыслями: как ставить задачу (здесь классификация бинарная), на какие данные и признаки обратить внимание. Обращают внимание на 3 типа данных: компания - например дата основания, количество раундов, область; инвесторы - имена, типы, сумма и т.д; выход компании.

В работе [4] про немного другой взгляд на инвестирование. Основной вывод - личность и опыт фаундера(ов) имеет очень большое влияние на успех компании.
В статье [5] Два вида предсказаний (классификация). Первый - успешный выход на IPO или приобретение, банкротство или что останутся частными. То есть на основании фичей, мы определяем, к какому классу относится компания, учитывая ее текущее состояние. Второй - получит ли доступ к дополнительному финансированию. Бинарная классификация. Говорят что модель уже развернута на AWS, и позволяет делать рилтайм запросы и предсказания. Используют ансамбль Deep Learning, XGBoost, Random Forests и K-Nearest Neighbors. Некоторые из фичей - количество раундов финансирования, локацию, количество, этап и дату раундов финансирования, а также инвестированную сумму вместе с должностями сотрудников и учредителей. Говорят что хотя можно представить историю стартапа с помощью временных рядов, основанных на последовательности раундов финансирования, вместо этого решили использовать одну фиксированную запись для каждого стартапа. (Тут скажу, что к задаче оттока пользователей ровно такие же два подхода, и зачастую TS подход фэйлится). Это достигается путем аггрегирования фичей - среднее время между раундами, сумма финансирования, количества раундов и информация о последнем раунде финансирования.
Однин из интересных подходов анализа данных описан в работе [6]. Помимо данных о компаниях они использовали LinkedIn о фаундерах.
Еще одна интересная статья [7] чем то схожая с [5]. Важное - рассматривают компании на ранней стадии. То есть с небольшим возрастом и находящиеся на ранней стадии финансирования (ранее, чем серия С).Берут все таки время в рассмотрение. Они пытаются предсказать, какое следующее действие произойдет с компанией в заданном временном окне. То есть тоже сводят к классификации. Классы - компания закрывается; получает инвестиции через новый раунд финансирования; компания приобретается другой компанией; или она становится публичной через IPO (что позволяет компании привлекать капитал на публичных рынках). Из статьи “Мы также рассматриваем случай, когда для компании в Crunchbase не сообщается ни о каком событии во время окна моделирования.” - добавляют класс NO EVENT, то есть ни один из перечисленных. В общем их идея в чем - берут на время какого то среза молодую компанию (по их филтру), к моменту среза собирают все фичи, и в следующем временном окне пытаются предсказать, что с ней случилось. В статье кстати тоже подняли наблюдение про малое количество фейлов - “Компании, добившиеся успеха, в некоторой степени перепредставлены, потому что те, которые потерпели неудачу на раннем этапе, возможно, вообще не фигурировали в базе данных.”
Статья [8] интересна тем, что рассматривается Look-ahead bias-free подход.
Работа [9] еще один пример анализа компаний на ранней стадии, когда нет большого колиества информации о компании и венчурные капиталисты часто полагаются на интуицию или эвристику при принятии решения, которое является предвзятым и потенциально вредным. Также рассматривается таксономия сигналов, способных предсказать вероятность успеха стартапа. Интересно.

4 Метрики качества

Качество обученных моделей будем определять по метрике Accuracy и ROC-curve. Ожидается, что их значения будут больше 85%.

5 Примерный план

- 23.04 подготовить чистые данные
- 30.04 обучить модельки, сделать черновой вариант постер в latex
- 07.05 дописать постер
- Оформить UI для будущего web-приложения, если будет время

References

[1] Kyle Wiggers. Gartner: 75% of vcs will use ai to make investment decisions by 2025. *VentureBeat*, 2021.

[2] Charlotta Siren Dietmar Grichnik Malin Malmstrom Torben Antretter, Ivo Blohm and Joakim Wincent. Do algorithms make better — and fairer — investments than angel investors? *Harward Business Review*, 2020.

[3] Wesley Klock. Picking winnig startups. *Medium*, 2018.

[4] Olga Maslikhova. Momentum vs. value-based investing in venture capital: The power of conviction. *Medium*, 2021.

[5] Greg Ross, Sanjiv Das, Daniel Sciro, and Hussain Raza. Capitalvx: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7:94–114, 2021.

[6] David Scott Hunter, Ajay Saini, and Tauhid Zaman. Picking winners: A data driven approach to evaluating the quality of startup companies, 2018.

[7] Javier Arroyo, Francesco Corea, Guillermo Jimenez-Diaz, and Juan A. Recio-Garcia. Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, 7:124233–124243, 2019.

[8] Kamil Żbikowski and Piotr Antosiuk. A machine learning, bias-free approach for predicting business success using crunchbase data. *Information Processing Management*, 58(4):102555, 2021.

[9] Francesco Corea, Giorgio Bertinetti, and Enrico Maria Cervellati. Hacking the venture industry: An early-stage startups investment framework for data-driven investors. *Machine Learning with Applications*, 5:100062, 2021.