Мастерская 1

Предсказание статуса стартапа (работает/закрыт)

www.kaggle.com/competitions/startups-predictions-m130ds/overview

Содержание

- 1. Какая стояла задача
- 2. Какие были вводные
- 3. Как решали (какие методы, инструменты применили)
- 4. Какие сложности встретили
- 5. Как решали проблемы
- 6. Какое решение получилось

Задача: Предсказание статуса стартапа

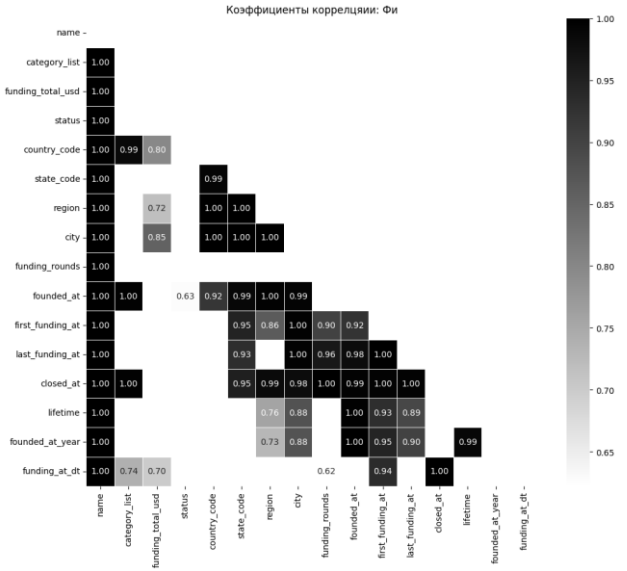
- Предсказать какие стартапы, функционирующие в период с 1970 по 2018 годы, закроются.
- Метрика: F1.
- Минимальное значение метрики: 0,71186.
- Значение метрики на случайной модели: 0,15270.
- Ссылка: www.kaggle.com/competitions/startups-predictions-m130ds/overview

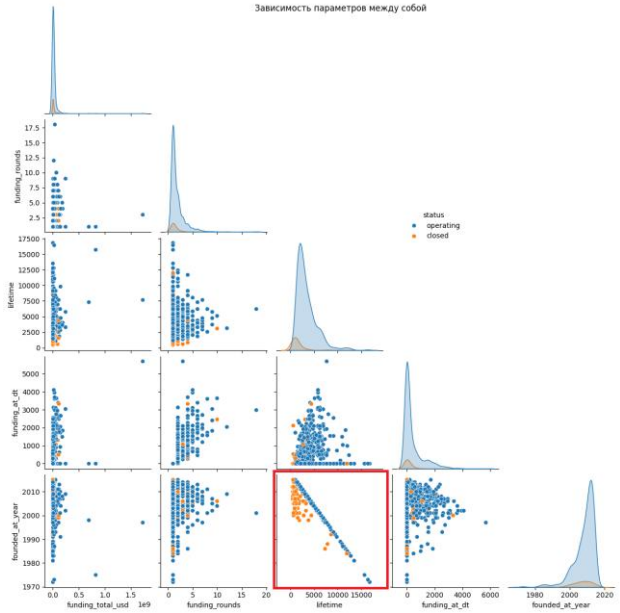
Какие были вводные

- Получить максимальное значение метрики F;
- Использовать библиотеки и технологии: pipeline, imblear, sklearn, Catboost, lightgbm; phik; shap;
- Сформулировать рекомендации позволяющие повысить шанс на успех стартапа;
- Подготовить отчет по исследованию.

Как решали: исследование данных

- Построены различные представления данных (~31):
  - общая статистика;
  - количество стартапов по категориям;
  - количество стартапах по странам.
- Построена визуализация зависимостей
- Рассчитана корреляция
- Сформированы новые параметры (~3):
  - Время существование стартапа
  - Год стартапа
  - Длительность инвестиций.

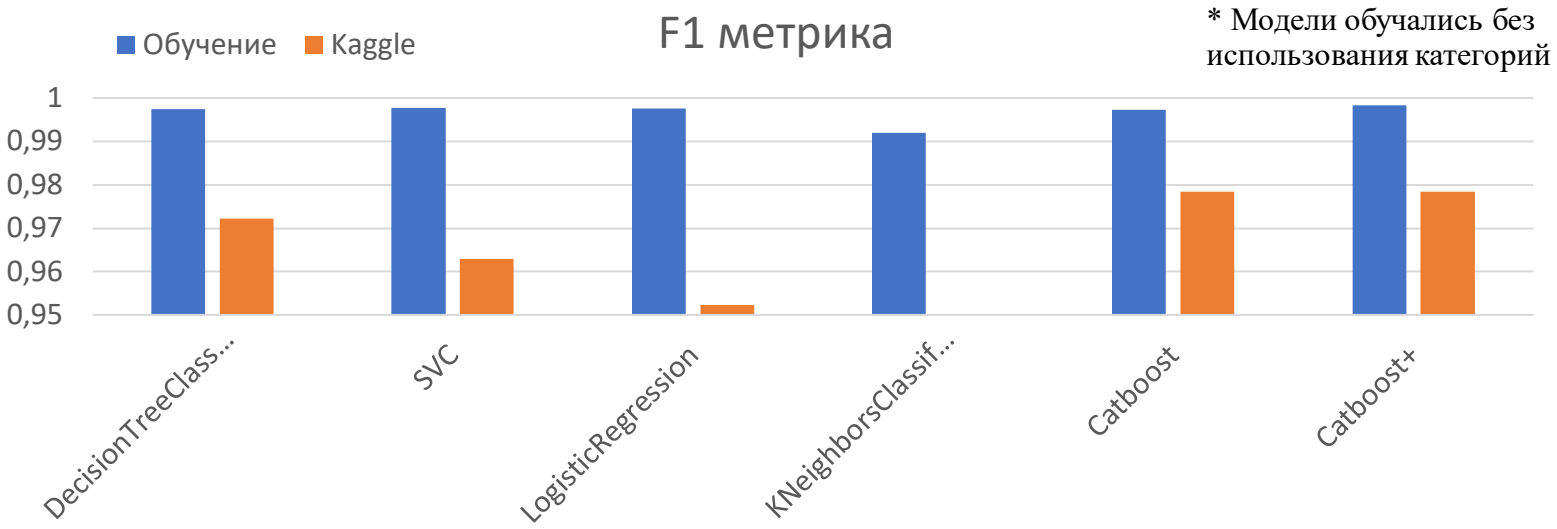




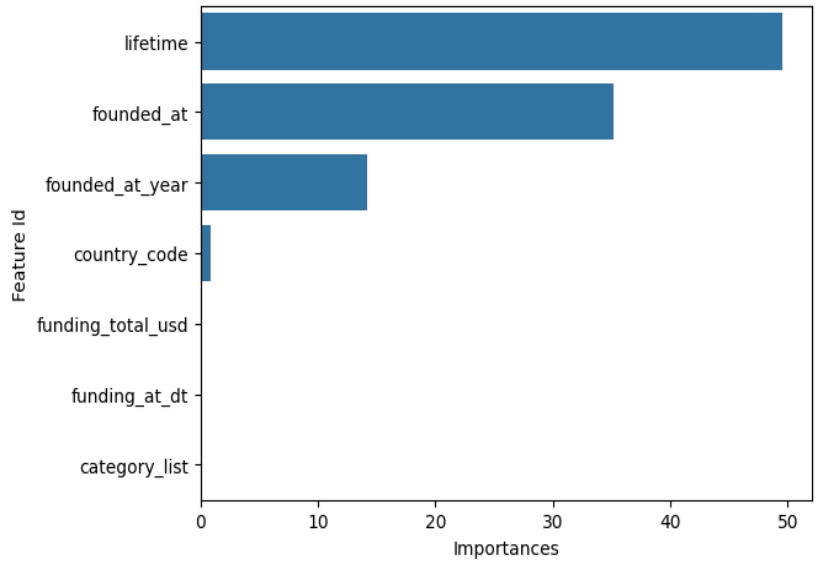
Название стартапа (name)	Список категорий, к которым относится стартап (category_list)	Общая сумма финансирования в USD (funding_total_usd)	Статус стартапа (закрыт или действующий) (status)	Код страны (country_code)
Код штата (state_code)	Регион (region)	Город (city)	Количество раундов финансирования (funding_rounds)	Дата основания (founded_at)
Дата первого раунда финансирования (first_funding_at)	Дата последнего раунда финансирования (last_funding_at)	Дата закрытия стартапа (если применимо) (closed_at)	Время существования стартапа в днях (lifetime)	

Как решали: Модели классификации

- Рассмотрены различные модели классификации:
  - DecisionTreeClassifier;
  - SVC;
  - LogisticRegression;
  - KNeighborsClassifier ;
  - Catboost.
- Для лучшей модели рассмотрены варианты обучения:
  - с расширенным набором данных за счет увеличения категорий;
  - с первичным набором данных.



* Модели обучались без использования категорий



Какие сложности встретили

- Выявление зависимостей и формулирование гипотез;
- Соблюдение баланса между высоким значением метрики и использования свойств компании;
- Улучшение модели за счет входных данных.

Как решали проблемы

Выявление зависимостей и формулирование гипотез	Соблюдение баланса между высоким значением метрики и использования свойств компании	Улучшение модели за счет входных данных.
• Рассмотрены ~31 варианта представления данных; • Сформулирована гипотеза, что данные не зависят от страны, но из-за небольшого количества данных имеются смещения;	• Приоритет был отдал получению высоких значений метрик; • Использование свойств компании вынесено на дальнейшее развитие;	• Увеличение количества исходных данных

Какое решение получилось

- Все модели показывали примерно одинаковый результат;
- Финальным решением было выбрана модель catboost;
- Реализовано два сценария обучения:
  - без обогащения данных (F1 метрика на обучающей - 0,997);
  - с обогащением данных (F1 метрика на обучающей - 0,998).

Дальнейшие шаги для исследования:

- исключить зависимость с параметром lifetime. Например, разбить данные на несколько строк с разной длительностью;
- убрать не явные дублирования в данных категорий;
- использовать методы борьбы с дисбалансом.

Замечания

- Тестовые данные могут изменяться. Например, можно заполнять пропуски нулями, удалять пропуски, считать производные фичи (для которых используются уже имеющиеся и т.д.). Главное правило – мы не можем использовать те данные, которые неоткуда взять;
- Длительность стартапа надо использовать с осторожностью или не использовать вовсе. Дело в том, что он ничего не говорит о том, что это за компания, чем она занимается и т.д. И есть опасность просто предсказывать компании, которые достаточно долго существуют, чтобы стать устойчивыми;
- Сравнение модели с dummy. В силу дисбаланса классов при определенной стратегии он способен был бы показать метрику больше 90%;
- Интересным было бы исследование, основанное только на свойствах компаний, а не на временных фичах.