нетология

Пояснительная записка
к дипломной работе на тему:

"Машинное обучение для предсказания дефолта по кредиту"

Автор: Малинина Елена
Группа: DSU - 1

Оглавление

1. Постановка задачи для машинного обучения

Машинное обучение (Machine Learning, ML) - форма искусственного интеллекта (ИИ), которая позволяет системе итеративно обучаться на данных, используя различные алгоритмы для описания данных и прогнозирования результатов.

В данной работе рассматривается метод прогнозирование кредитоспособности клиентов банка Home Credit. Исходные данные получены с сайта www.kaggle.com . Сам датасет можно посмотреть по ссылке.

Кредитный скоринг представляет собой математическую или статистическую модель, с помощью которой на основе кредитной истории «прошлых» клиентов банк пытается определить, насколько велика вероятность, что конкретный потенциальный заемщик вернет кредит в срок.

Повышение доходности кредитных операций непосредственно связано с качеством оценки кредитного риска. В зависимости от классификации клиента по группам риска банк принимает решение, стоит ли выдавать кредит или нет, какой лимит кредитования и проценты следует устанавливать.

Основная задача – **оценка риска**

Целевая метрика следующая:

- TP - истинно-положительное решение
- TN - истинно-отрицательное решение
- FP - ложно-положительное решение
- FN - ложно-отрицательное решение

1. Accuracy - описывает общую точность предсказания модели по всем классам.

$$\text{Accuracy} = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + TN + FP}$$

2. Precesion - отношение TP к TP + FP. Это доля объектов, названными классификатором положительными и при этом действительно являющимися положительными

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. Recall - отношение TP к TP + FN. Это то, какую долю объектов положительного класса из всех объектов положительного класса нашёл алгоритм

$$\text{Recall} = \frac{TP}{TP+FN}$$

4. F-мера - представляет собой гармоническое среднее между точностью и полнотой.Максимальный F1-score мы получим, если и recall, и precision достаточно далеки от нуля.

$$F1 = 2 * \frac{recall * precision}{recall + precision}$$

2. Анализ данных и поиск аномалий

В анализ данных был взять только датасет application_train, так как датасет applicatin_test содержит только тестовые данные.
Основной датасет application_train содержит 122 колонки, из них 16 колонок содержат категориальные переменные. Целевая переменная (target) несбалансированна, имеется большой перевес в сторону 0-го класса, где 0 - кредит погашен, 1 - кредит не погашен. При обучении модели это приведет к неверному прогнозу в сторону 0-го класса. Взяты также несколько дополнительных датасетов:

- previous_application - все предыдущие заявки на кредит
- POS_CASH_balance.csv - баланс на кредитном счете
- credit_card_balance.csv - баланс на предыдущем кредитном счете

Общий датасет составил 207 колонок.

Если взглянуть на колонку с типом кредитования, можно заметить, что возвратные кредиты пользуются меньшей популярностью, соответственно и проблем с их выплатой меньше.

Далее рассматривается пол заемщика и чаще всего это женщины. Судя по графику проблемы с выплатой у женщин почти наравне с мужчинами, хотя мужчины берут кредиты реже. Также присутствует неизвестный класс XNA, значений в нем всего 4 шт. и их можно убрать.

Возраст клиента тоже представляет интерес. Разброс возраста от 20 до 70 лет, всплеск кредитования приходится на 25 - 45 лет. На графике ниже можно увидеть, что проблемы с выплатой приходится примерно на тот же период жизни. При этом, чем старше человек, тем меньше вероятность не погасить кредит. На boxplot можно увидеть, что выбросов нет.

Далее обратим внимание на возраст машины клиента. Есть значения, которые превышают отметку в 65 лет. Их немного и лучше их убрать, они могут исказить прогноз.

Семейный статус клиента может быть весомым при прогнозе, его тоже рассмотрим. Большинство женаты / замужем. Есть класс Unknown, там представлены 2 строки, его убираем.

Взглянем на количество детей, большинство детей не имеют. На boxplot есть отметки, где больше 10 детей. Этих данных всего 8 строк, их тоже уберем, так как может быть перекос в данных.

Далее на графике типа занятости можно увидеть, что большая часть людей имеют работу, но так же много и пенсионеров. В boxplot годовом доходе заемщика есть значения превышающие 12 000 000. Эти данные лучше убрать из - за большого разрыва.

Далее представлен график суммы кредита и он очень сильно варьируется. Самые большие кредиты от 4 000 000 брали коммерческие партнеры, данные выглядят вполне адекватно. Посмотрим на boxplot, где выражено отношение профессии и суммы кредита, увидим значительные выбросы. Их в общем около 15 штук и их отбросим.

Далее взглянем на количество отработанных дней. Есть значение в 365243 дня, оно равняется 1000 лет. Эти данные составляют 18% от всего дата фрейма, поэтому эти значения заменяем на нули.

В матрицах корреляции отберем только несколько значений, которые меньше всего связаны с таргетной переменной (FLAG_DOCUMENT_6, FLAG_DOCUMENT_13, FLAG_DOCUMENT_16, FLAG_DOCUMENT_14, FLAG_DOCUMENT_20).

Далее используем Label Encoder для кодирования категориальных переменных и заполняем нули модой с помощью библиотеки Simple Imputer.

Опираясь на матрицу корреляций, создаем новые переменные, чтобы разорвать связь:

- RATIO_CREDIT_PRICE_% - отношение кредита и стоимости покупок
- RATIO_CREDIT_ANNUITY_% - представление о сроках кредита
- RATIO_GOODS_INCOME_% - отношение зарплаты и стоимости покупок

3. Методика реализации

Для начала запишем целевое значение в отдельную переменную target. В данных присутствует большой разброс и нужно их нормализовать с помощью MinMaxScaler в диапазоне от 0 до 1. Далее используется библиотека SMOTE для устранения дисбаланса в target. Т.к. минорный класс очень задавлен, будем дублировать его по 100 ближайшим соседям. После этого разделим данные на тренировочную и тестовую часть и начнем обучение логистической регрессии.
После этого обучим модель градиентного бустинга, где используется ансамбль слабых моделей деревьев решений, CatBoostClassifier с параметрами:

- 500 итераций
- детектор переобучения - 20
- скорость обучения - 0.01
- глубиной дерева - 7
- применение жадного алгоритма grow_policy, где листья разделяются с использованием условия, которое приводит к лучшему уменьшению потерь

4. Итоги обучения

Модель логистической регрессии не смогла показать хороший результат в решении этой задачи, точность в районе 70%.

Точность в 95% показывает, что модель CatBoostClassifier хорошо подошла для реализации кредитного скоринга.

5. Выводы и заключение

Полученные результаты демонстрируют, что данная задача успешно решается методами машинного обучения, и может достигать высокой точности, так же открываются новые возможности для анализа и оптимизации оценки риска.
В данной работе был произведен анализ исходных данных, выявлены и устранены аномалии. Произведено сравнение результатов модели логистической регрессии и градиентного бустинга. Создание модели машинного обучения способной оценить риск при кредитовании клиента. Возможно, добавление новых признаков (как анализ соц.сетей, например) повысит качество моделей.