Задание проекта по кредитному риску

Каждая команда может выбрать задание либо по кредитному скорингу физических лиц, либо по кредитному риску компаний. Задания примерно похожи по трудоёмкости и состоят из схожих этапов, но отличаются расстановкой акцентов и сравнительной трудоёмкостью этапов.

Задание по кредитному скорингу

Построить модель предсказания вероятности дефолта заёмщика – физического лица по данным, доступным в заявке.

Главное — результат и его успешная коммуникация. Использовать WoE и классический подход к скорингу не обязательно (хотя это может помочь в коммуникации 😉). Для работы с WoE можно использовать пакет: https://github.com/sberbank-ai/wing или любой другой.

Финальным результатом работы считаются ответы на вопросы 6 или 7. В случае ошибок в промежуточных расчётах существенными будут считаться ошибки, сильно повлиявшие на ответы на вопросы 6 и 7, а несущественными — ошибки, не сильно повлиявшие на эти ответы.

Коммуникация результата считается успешной, если вас поняли на защите без 100500 дополнительных вопросов 😉.

Описание данных

Мы будем использовать данные Lending Club, доступные как общественное достояние (CC0) по следующей ссылке: https://www.kaggle.com/wordsforthewise/lending-club

Внимание! Это достаточно сырые данные. В них возможны, хотя и не очень вероятны, ошибки. Эти данные собраны постфактум, так что некоторые поля некорректно использовать в скоринге, так как их значения недоступны на этапе рассмотрения заявки.

Задание

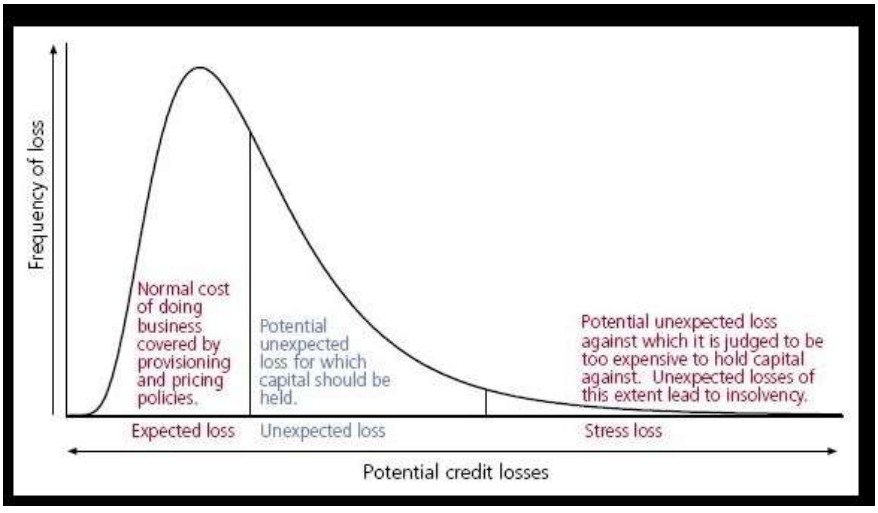
1. Разберитесь в характере доступных данных. Оставьте для дальнейшего использования только те признаки, которые доступны при заявке. Убедитесь, что вы хорошо понимаете смысл всех полей. Сформируйте набор признаков и целевую переменную. Не забудьте обосновать свой выбор. Если в дальнейшем обнаружится «утечка данных» из-за выбранного признака, это будет очень серьёзной ошибкой, так как обесценит практически всю последующую работу.
2. Разбейте выборку на несколько частей для построения модели и анализа качества. Осторожно! Ошибки в разбиении выборки могут обесценить значительную часть результатов (например, валидация in-sample практически бесполезна).
3. Постройте интерпретируемый скоринг. Для этого:
    a. Проведите предварительный анализ данных.
    b. Отберите переменные для анализа, руководствуясь предварительным анализом и здравым смыслом.
    c. Постройте интерпретируемую модель и оцените её качество.
    d. Интерпретируйте оценённую зависимость в терминах влияния отдельных признаков и их комбинаций, если они вошли в модель, на вероятность дефолта.
4. Постройте любую модель машинного обучения на этих же данных с качеством выше, чем у построенной ранее интерпретируемой модели. Кратко опишите процесс построения и результаты.
5. Проведите валидацию обеих построенных моделей. Вы можете выбирать тесты для валидации самостоятельно, руководствуясь материалами курса, научными и методическими работами,

нормативными документами ЦБ и других национальных регуляторов, документами Базельского комитета и пр. Необходимо раскрыть, по меньшей мере, следующие грани (на каждую грань можно делать более одного теста, если это нужно для полноты картины).

a. Некоторые из тестов подразумевают дискретизацию предсказаний (внутренние рейтинги, основанные на модели). Проведите эту дискретизацию один раз и в дальнейшем используйте одно и то же разбиение непрерывного скорингового балла на внутренние рейтинги во всех тестах.
b. Разделяющая способность модели. Продемонстрируйте, что модель хорошо отделяет дефолтёров от не-дефолтёров.
c. Переобучение. Нет ли признаков переобучения модели?
d. Простота модели. Нельзя ли упростить модель, не сильно потеряв при этом в качестве? Действительно ли нужны все эти факторы и все эти взаимосвязи?
e. Устойчивость модели во времени. Продемонстрируйте, что модель стабильно ранжирует наблюдения на всех временных срезах и отсутствуют периоды, в которых разделяющая способность модели недостаточна.
f. Если в модели использовалась дискретизация переменных, оцените экономический смысл разбиения переменных на именно такие категории. Нет ли необъяснимых немонотонностей и/или неустойчивостей?
g. Корректность оценки PD. Проверьте, согласуются ли предсказания PD и реально наблюдаемая дефолтность. Если нет, это не страшно. Разработайте поправочную модель и продемонстрируйте, что она выполняет свою функцию.

6. Пусть в портфеле находятся все кредиты, выданные 01.01.2017 или позже, притом без учёта признака дефолта — для целей этого упражнения мы предполагаем, что они все активны в настоящее время. Предположим, что LGD=100%, а текущая задолженность по каждому кредиту равна полю *funded_amnt*. При помощи обеих моделей рассчитайте ожидаемые потери (математическое ожидание) и неожидаемые потери (VaR на уровне 99.9%). Рассчитайте необходимый капитал как разность этих двух величин. Не забудьте доверительные интервалы.
   a. В предположении независимости дефолтов по отдельным кредитам.
   b. С учётом корреляций между дефолтами в 6%.

   Постройте такую иллюстрацию и отметьте на ней фактические потери по валидационной выборке. Форма распределения потерь на графике — для примера; у вас будет другая.



7. Принимая LGD=100%, рассчитайте для обеих моделей по обучающей выборке порог отсечения, максимизирующий ожидаемую прибыль от кредитования. Постройте графики зависимости ожидаемой прибыли от порога отсечения (для двух моделей в одних осях). Не забудьте доверительные интервалы.

Отчётные материалы

- Отдельно ответы на вопросы 6 и 7 — по каждой из двух построенных моделей. Эта информация будет использоваться нами в агрегированном виде. Мы поделимся статистикой после защит. Всего нужно отдельно сдать 6 ответов:
    - размер капитала для интерпретируемой модели и независимых дефолтов (число и 99% доверительный интервал);
        - то же самое для продвинутой модели;
    - размер капитала для интерпретируемой модели и коррелированных дефолтов (число и 99% доверительный интервал);
        - то же самое для продвинутой модели;
    - коэффициент Джини и максимальная прибыль (число и 99% доверительный интервал) по интерпретируемой модели;
        - то же самое для продвинутой модели.

- Расчётный файл / код.

  Требования: полная воспроизводимость результатов (фиксируйте random seed), возможность обращения к промежуточным результатам, user-friendly (комментарии в коде и/или краткое руководство пользования с описанием входов / преобразований данных / выходов). Если вы используете код, написанный не вами, обязательно указывайте автора и источник заимствования. Без этого заимствованный код будет считаться с плагиатом.

- Презентация для устной защиты. Регламент: 20 мин.

  На слайдах представить тезисы и иллюстративные материалы; текст слайдов должен дополнять/раскрывать устное выступление, но не дублировать его.

  Все использованные в презентации иллюстрации должны в точности генерироваться сданным кодом!

Задание по кредитным рейтингам

Построить модель оценки вероятности отзыва банковской лицензии по данным отчётности банка.

Данные

Мы будем использовать данные по отзыву банковских лицензий и банковской отчётности с сайта ЦБ РФ (http://cbr.ru/). Информация по кредитным рейтингам бесплатно доступна на сайте https://www.bankodrom.ru/. Вы можете либо распарсить эти сайты, либо найти эту же информацию в более удобном для машинной обработки виде в разнообразных базах данных, доступных по подписке ВШЭ (http://sophist.hse.ru/4dbank.shtml). При необходимости можно использовать дополнительную информацию.

Задание

1. Сформируйте целевую переменную. Обратите внимание на то, что причины отзыва банковской лицензии могут быть разными. Обратите внимание на формулировку как целевого события, так и горизонта.
2. Сформируйте необходимое для обучения число подвыборок. Обоснуйте как способ формирования общей выборки (как именно формируются строки в таблице, чтобы избежать утечки данных), так и способ разбиения на подвыборки. Остерегайтесь автоматического разбиения — ввиду сравнительно малого объёма выборки возможны статистические артефакты.
3. Постройте модель предсказания вероятности отзыва лицензии. Интерпретируемость здесь не обязательна, но при выборе структуры моделей учтите ограниченность данных. Для построения модели:
    a. Проведите предварительный анализ данных.
    b. Отберите переменные для анализа, руководствуясь предварительным анализом и здравым смыслом.
    c. Постройте модель и оцените её качество.
    d. По возможности, интерпретируйте оценённую зависимость в терминах влияния отдельных признаков и их комбинаций, если они вошли в модель, на вероятность дефолта.
4. Постройте аналогичную модель на основе исключительно кредитных рейтингов, присваиваемых банкам кредитными рейтинговыми агентствами. Для целей задания достаточно использовать рейтинги одного кредитного рейтингового агентства, однако для улучшения результатов вы можете использовать рейтинги нескольких агентств — не забудьте описать, как именно вы это сделали. Кратко опишите результаты.
5. Проведите валидацию обеих построенных моделей, сравнивая их друг с другом. Вы можете выбирать тесты для валидации самостоятельно, руководствуясь материалами курса, научными и методическими работами, нормативными документами ЦБ и других национальных регуляторов, документами Базельского комитета и пр. Необходимо раскрыть, по меньшей мере, следующие грани (на каждую грань можно делать более одного теста, если это нужно для полноты картины).
    a. Некоторые из тестов подразумевают дискретизацию предсказаний (внутренние рейтинги, основанные на модели). Проведите эту дискретизацию один раз и в дальнейшем используйте одно и то же разбиение непрерывного скорингового балла на внутренние рейтинги во всех тестах.
    b. Разделяющая способность модели. Продемонстрируйте, что модель хорошо отделяет банки, у которых отозвали лицензию, от сохранивших её.

c. Переобучение. Нет ли признаков переобучения модели?
d. Простота модели. Нельзя ли упростить модель, не сильно потеряв при этом в качестве? Действительно ли нужны все эти факторы и все эти взаимосвязи?
e. Устойчивость модели во времени. Продемонстрируйте, что модель стабильно ранжирует наблюдения на всех временных срезах и отсутствуют периоды, в которых разделяющая способность модели недостаточна.
f. Если в модели использовалась дискретизация переменных, оцените экономический смысл разбиения переменных на именно такие категории. Нет ли необъяснимых немонотонностей и/или неустойчивостей?
g. Корректность оценки вероятности отзыва. Проверьте, согласуются ли предсказания вероятности отзыва и реально наблюдаемая частота отзывов. Если нет, это не страшно. Разработайте поправочную модель и продемонстрируйте, что она выполняет свою функцию.

6. Пусть на сегодняшний момент у нас есть экспозиция к банкам «Альфа-Банк», «Авангард», «Росбанк», «Экспобанк» и «Локо-Банк» — по 200 млн. руб. к каждому. Мы не ожидаем изменения этой суммы. Примем LGD = 40%. Используя обе свои модели, оцените:
    a. Ожидаемые потери (математическое ожидание) и неожидаемые потери (VaR на уровне 99.9%) на горизонте 1 год. Оцените экономический капитал как разность этих двух величин.
    b. Проведите аналогичный расчёт для горизонта 5 лет. Учтите, что на горизонте 5 лет влияние изменения макроэкономических показателей и/или бизнес-циклов может быть ощутимым.
7. Оценка по какой из моделей (на основе кредитных рейтингов или на основе отчётности) вам видится более адекватной? Аргументируйте выбор в первую очередь результатами валидации обеих моделей с использованием дополнительной информации по желанию.

Отчётные материалы

- Отдельно ответы на вопрос 6 — по каждой из двух построенных моделей. Эта информация будет использоваться нами в агрегированном виде. Мы поделимся статистикой после защит. Всего нужно отдельно сдать:
    - ожидаемые потери и размер капитала для модели по отчётности (число и 99% доверительный интервал);
    - то же самое для модели по кредитным рейтингам.

- Расчётный файл / код

  Требования: полная воспроизводимость результатов (фиксируйте random seed), возможность обращения к промежуточным результатам, user-friendly (комментарии в коде и/или краткое руководство пользования с описанием входов / преобразований данных / выходов). Если вы используете код, написанный не вами, обязательно указывайте автора и источник заимствования. Без этого заимствованный код будет считаться с плагиатом.

- Презентация для устной защиты. Регламент: 20 мин.

  На слайдах представить тезисы и иллюстративные материалы; текст слайдов должен дополнять/раскрывать устное выступление, но не дублировать его.

Все использованные в презентации иллюстрации должны в точности генерироваться сданным кодом!