### Оценка модели CCF с использованием t-test

#### Введение

Оценка модели кредитного конверсии (CCF, Credit Conversion Factor) с помощью t-теста позволяет проверить, насколько значимо отличаются прогнозируемые моделью значения CCF от фактических значений. Этот тест применяется как для анализа всего портфеля, так и для отдельных сегментов, что позволяет выявить, в каких группах модель работает недостаточно точно.

#### Постановка задачи

t-тест используется для проверки нулевой гипотезы, которая утверждает, что среднее значение прогнозируемых CCF () совпадает с фактическим средним значением CCF (). Формально:

Альтернативная гипотеза:

#### Методика проведения t-test

1. **Подготовка данных**
   * Выберите выборки из данных, где имеются прогнозируемые () и фактические () значения.
   * Убедитесь, что данные очищены от выбросов, пропусков и ошибок, которые могут исказить результаты анализа.
2. **Расчет t-статистики**  
   t-тест для независимых выборок рассчитывается по формуле:

* где:
  + , : средние значения прогнозируемых и фактических CCF;
  + , : дисперсии прогнозируемых и фактических значений;
  + , : объемы выборок для прогнозируемых и фактических значений.

1. **Оценка p-value**  
   На основе рассчитанной t-статистики и степеней свободы () определяется p-value. Степени свободы вычисляются как:
2. **Принятие решения**
   * Если -value меньше уровня значимости (), то нулевая гипотеза отклоняется, и считается, что средние значения CCF существенно различаются.
   * Уровень значимости обычно выбирается равным 0.05, но может быть уменьшен (например, до 0.01) для повышения строгости теста.

#### Применение t-теста к сегментам

Для повышения точности модели полезно проверить гипотезу отдельно по различным сегментам портфеля. Например:

* **Сегменты заемщиков:** корпоративные клиенты, физические лица.
* **Сегменты по типам продуктов:** кредитные карты, овердрафты, кредитные линии.
* **Сегменты по географии или отрасли.**

1. Разделите данные на подгруппы в соответствии с выбранными сегментами.
2. Проведите t-тест для каждой подгруппы. Это позволит выявить, в каких сегментах модель работает лучше или хуже.

#### Интерпретация результатов

* Если -value значимо меньше , это указывает на то, что модель требует доработки, так как она недостаточно точно прогнозирует CCF.
* Если -value больше , это говорит о том, что средние прогнозы и фактические значения статистически совпадают.

#### Преимущества t-теста

1. Простота расчета и интерпретации.
2. Возможность применения как для всего портфеля, так и для отдельных сегментов.
3. Статистическая строгость, позволяющая сделать обоснованные выводы о точности модели.

#### Ограничения t-теста

1. Требование нормальности распределения данных (для выборок малого объема).
   * Если выборки велики (), можно полагаться на центральную предельную теорему.
2. Чувствительность к выбросам.
   * Выбросы могут существенно исказить результаты анализа, поэтому важно заранее проверять данные.

Таким образом, t-тест является важным инструментом для проверки качества прогнозов модели CCF. Его результаты позволяют выявить сегменты портфеля, где требуется улучшение модели, что особенно важно для управления кредитными рисками.

### Проверка корректности сегментации модели CCF

#### Введение

Сегментация играет ключевую роль в построении и оценке моделей, таких как модель CCF. Она позволяет адаптировать модель к особенностям различных групп заемщиков или продуктов, улучшая точность прогнозирования. Проверка корректности сегментации направлена на подтверждение того, что выбранные сегменты являются логически обоснованными и статистически различимыми, а также что они вносят вклад в повышение качества модели.

#### Цели проверки корректности сегментации

1. **Оценка различий между сегментами:** Проверка, действительно ли сегменты статистически отличаются друг от друга по ключевым характеристикам, таким как средние значения CCF, уровень риска, тип продукта и т. д.
2. **Подтверждение эффективности сегментации:** Анализ, приводит ли сегментация к улучшению точности модели.
3. **Выявление избыточной или некорректной сегментации:** Проверка, не создает ли сегментация избыточное разделение данных, которое не повышает качество модели или даже ухудшает ее производительность.

#### Методика проверки корректности сегментации

1. **Описание сегментов**  
   На первом этапе необходимо четко определить, какие критерии использовались для сегментации, например:
   * Тип продукта (кредитные линии, карты, овердрафты).
   * Размер компании (малый, средний, крупный бизнес).
   * Географическое положение.
   * Отрасль заемщика.
   * Кредитное качество (рейтинг, PD).

* Описание должно включать количественное распределение данных по сегментам, чтобы убедиться в их достаточной представленности.

1. **Статистическое обоснование сегментации**  
   Для проверки того, что сегменты действительно различаются, используются статистические тесты:
   * **ANOVA (дисперсионный анализ):**  
     Используется для сравнения средних значений CCF между сегментами.  
     Нулевая гипотеза (): Средние значения CCF для всех сегментов одинаковы.  
     Альтернативная гипотеза (): Средние значения CCF различаются хотя бы для одного сегмента.  
     Если p-value меньше уровня значимости (, обычно 0.05), то сегменты статистически различны.
   * Формула для F-статистики в ANOVA:
   * где:
     + : Средняя межгрупповая сумма квадратов.
     + : Средняя внутригрупповая сумма квадратов.
   * **t-test для парных сравнений:**  
     В случае небольшого количества сегментов можно использовать t-test для проверки значимости различий между каждой парой сегментов. Это особенно полезно, если ANOVA показывает значимость различий, но не указывает, какие именно группы отличаются.
2. **Анализ эффективности модели по сегментам**  
   Проверяется, насколько сегментация улучшает предсказательную способность модели:
   * Оцените метрики качества модели (например, RMSE, MAE) по каждому сегменту.
   * Сравните результаты модели на уровне сегментов с результатами по всему портфелю. Если сегментация приводит к снижению ошибок прогноза, она считается эффективной.
3. **Анализ однородности внутри сегментов**  
   Важно проверить, что данные внутри каждого сегмента однородны, а не содержат сильные различия. Для этого можно использовать:
   * **Коэффициент вариации:** Проверка разброса CCF внутри сегмента. Если вариативность велика, это может быть сигналом к необходимости более глубокой сегментации.
   * **Кластеры:** Использование алгоритмов кластеризации (например, k-means) для проверки, не содержит ли сегмент скрытых подгрупп.
4. **Проверка избыточной сегментации**  
   Используйте сравнение метрик модели на уровне сегментов и портфеля, чтобы убедиться, что сегментация не добавляет ненужной сложности. Избыточная сегментация может привести к переобучению модели или снижению ее интерпретируемости.

#### Пример проверки

* **Сегментация по типу продукта:** Проверьте, отличаются ли средние значения CCF для кредитных карт, кредитных линий и овердрафтов. Если различия статистически значимы, сегментация оправдана.
* **Сегментация по размеру компании:** Проверьте, одинаковы ли значения CCF для малых, средних и крупных компаний. Если различий нет, возможно, сегментация по этому критерию не нужна.

#### Интерпретация результатов

* Если сегменты статистически различимы, это подтверждает обоснованность сегментации.
* Если сегменты не отличаются или не приводят к улучшению качества модели, следует пересмотреть критерии сегментации.

#### Заключение

Проверка корректности сегментации модели CCF позволяет убедиться, что сегменты имеют логическую и статистическую обоснованность. Этот этап критически важен для создания надежной модели, которая эффективно отражает различия между группами заемщиков или продуктов, минимизируя ошибки прогнозирования.

### Бэктест модели CCF

#### Введение

Бэктест (backtesting) модели CCF позволяет оценить, насколько хорошо прогнозы модели соответствуют фактическим значениям в исторических данных. Это ключевой этап проверки модели, направленный на выявление систематических ошибок и оценку ее предсказательной способности.

#### Подход к проведению бэктеста

1. **Постановка задачи**  
   Цель бэктеста — сравнить прогнозируемые значения CCF () с фактическими значениями () для ретроспективного периода, используя метрики точности и статистические тесты.
2. **Выбор данных для бэктеста**
   * Используются данные, не участвовавшие в обучении модели (out-of-sample).
   * Данные должны включать фактические значения CCF, рассчитанные на основе реальной информации о конверсии неиспользованных лимитов в задолженность.
3. **Метрики для сравнения прогнозов и фактических значений**  
   Для количественной оценки точности модели используются следующие метрики:

#### Формулы для метрик

1. **Средняя абсолютная ошибка (Mean Absolute Error, MAE):**

* где:
  + : количество наблюдений;
  + : фактическое значение CCF для -го клиента;
  + : прогноз модели для -го клиента.

1. **Среднеквадратичная ошибка (Mean Squared Error, MSE):**
2. **Корень из среднеквадратичной ошибки (Root Mean Squared Error, RMSE):**

* RMSE более чувствителен к большим отклонениям, чем MAE.

1. **Коэффициент детерминации ():**

* где — среднее фактическое значение CCF. Значение близкое к 1, указывает на высокую точность модели.

1. **Средняя ошибка прогноза (Mean Forecast Error, MFE):**

* Используется для оценки смещения модели. Если значительно отклоняется от 0, модель имеет систематическую ошибку.

#### Графический анализ

1. **График ошибок:**  
   Постройте распределение ошибок () для выявления систематических смещений.
2. **График фактических и прогнозных значений:**  
   Постройте диаграмму рассеяния, где на оси отображаются прогнозные значения (), а на оси — фактические (). Линия указывает на идеальное соответствие.

#### Шаги проведения бэктеста

1. **Расчет метрик на уровне портфеля:**  
   Вычислите описанные метрики (MAE, RMSE, , MFE) для всего портфеля.
2. **Анализ по сегментам:**  
   Проведите те же расчеты для отдельных сегментов (по типу заемщика, продуктам, географии и т. д.). Это позволяет выявить сегменты, где модель недостаточно точна.
3. **Сравнение прогнозов и фактических значений:**  
   Постройте графики, отражающие соответствие прогнозов и фактических значений.
4. **Интерпретация результатов:**
   * Высокие значения RMSE или значительное смещение (MFE) указывают на необходимость доработки модели.
   * Низкие значения (например, ниже 0.5) свидетельствуют о слабой объясняющей способности модели.

#### Пример

Допустим, у вас есть следующие данные для 5 наблюдений:  
Фактические значения   
Прогнозируемые значения

Вычислим метрики:

#### Заключение

Бэктест модели CCF позволяет оценить точность прогнозов и выявить слабые места модели. Расчет количественных метрик и визуализация результатов дают объективное понимание качества модели, а также помогают принять решения о необходимости доработок.