МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»**

*ИНСТИТУТ* ИТАСУ

*КАФЕДРА АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ*

*НАПРАВЛЕНИЕ* 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

**Лабораторная работа №5**

**По дисциплине**: Технологии интеллектуального анализа данных

**На тему:** Метрики оценки качества модели классификации на примере задачи кредитного скорринга

**Студент: Костромин Д.А.**

**Группа: МИВТ 18-5-7**

**Преподаватель: Фомичева О.Е.**

**Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Дата защиты: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Москва, 2019

**Оглавление**

[**Введение** 3](#_Toc25025413)

[**1. Типы ошибок и метрики качества модели для задачи классификации** 4](#_Toc25025414)

[**2. Обработка исходных данных с помощью логистической регрессии и ансамбля деревьев решений** 7](#_Toc25025415)

[**3. Применение сэмплинга с целью корректировки баланса классов** 9](#_Toc25025416)

[**Выводы** 11](#_Toc25025417)

[**Список источников** 12](#_Toc25025418)

# **Введение**

В работе рассматриваются метрики, применимые для оценки качества классификационных моделей. Рассматривается задача бинарной классификации на примере кредитного скоринга, в том числе затрагивается применения “oversampling” с целью повышения качества модели при несбалансированной классификации.

При выполнении работы будут использованы модели машинного обучения из библиотеки scikit-learn для языка Python.

# **1. Типы ошибок и метрики качества модели для задачи классификации**

В задачах классификации обобщенная правильность не всегда является однозначным критерием корректности модели. В частности, возможно возникновение различных ошибок.

Вернемся к примеру с кредитным скорингом, автоматизированная система, основанная на машинном обучении, может пометить клиента как надежного (отрицательный класс, 0) или ненадежного, с высоким риском невозврата займа/наличием просрочек выплаты в кредитной истории (положительный класс, 1).

В таком случае возможна ошибка, при которой надежный клиент будет отнесен системой к классу ненадежных, подобное срабатывание системы называют ложно положительным (False Positive, FP) или “ошибка первого рода”. Аналогично возможно отнесение ненадежного клиента к классу надежных, в таком случае срабатывание системы можно охарактеризовать как ложно отрицательное (False Negative, FN) или “ошибка второго рода”. Типы ошибок при классификации принято визуализировать в виде матрицы ошибок, на главной диагонали которой, в задаче бинарной классификации, находятся правильно классифицированные объекты, истинно положительные (True Positive, TP) – клиенты, определенные как надежные и в действительности такими являющиеся и истинно отрицательные (True Negative, TP) – клиенты, классифицированные как ненадежные и соответственно являющиеся ненадёжными. Соответственно, ошибочно определенные объекты находятся вне главной диагонали матрицы, Рисунок 1. [1]

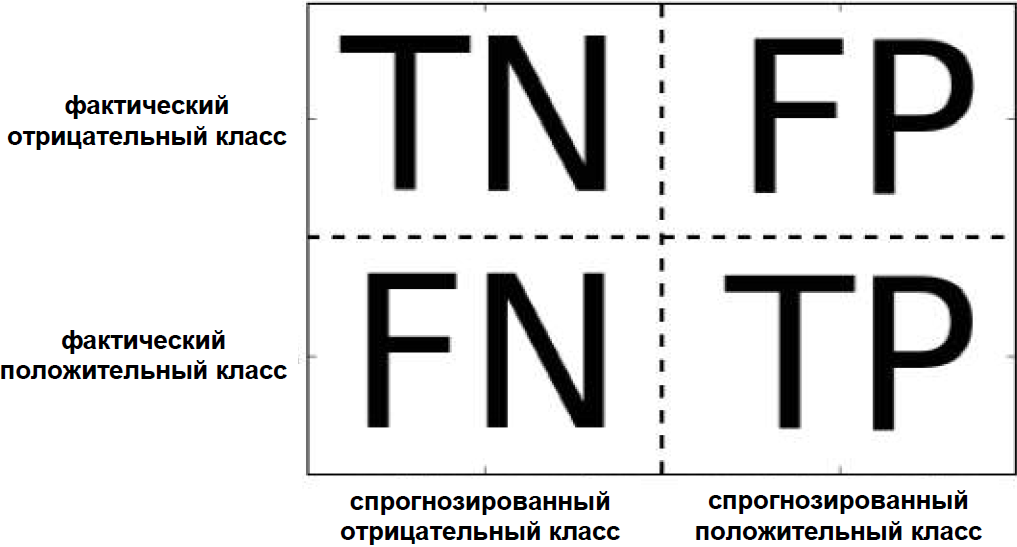


Рисунок 1 Типы ошибок при бинарной классификации

В работе рассматриваются следующие метрики классификации:

* Accuracy
* Precision
* Recall
* F-measure

Accuracy – Наиболее простая из метрик, является отношением количества правильных ответов к числу неправильных. Например, если верно классифицированы 9 объектов из 10, то значение accuracy будет равно 90% или 9/10. Главный недостаток accuracy – отсутствие оценки точности по каждому из классов объектов, например если решается задача бинарной классификации и в тестовой выборке имеется 95 объектов первого типа и 5 другого, и модель отнесла все 100 объектов к первому типу, то данная метрика покажет точность 95%, что может создать обманчивое впечатление о высоком качестве некорректной модели, относящей все объекты к одному классу. В библиотеке sklearn accuracy представлена функцией sklearn.metrics.accuracy\_score.

Precision – “Точность” модели, является отношением истинно-положительных ответов модели (TP) к суммарному числу положительных ответом модели, включая ложно-положительный (TP + FP). На оценку precision в первую очередь ориентируются в ситуации, когда необходимо минимизировать число ложно положительных ответов модели. В примере с кредитным скорингом, precision актуальна, если необходимо свести к минимуму число надежных клиентов, ошибочно классифицируемых как ненадежные. Но подобный подход опасен, так как не учитывается клиенты, ошибочно отнесенные к классу надежных.

Recall – “Полнота”, отношения количества истинно положительных (TP) объектов к общему числу объектов, классифицированных как положительные, включая ошибку второго рода – объектов, ошибочно определенных как положительные (TP + FN). Подобная метрика лучше подходит к для задачи кредитного скоринга, нежели предыдущие. Особенно актуальной recall становится, если бизнес-задача требует минимизацию ошибки второго рода – важно минимизировать количество ненадежных клиентов, даже если в результате увеличится количество надежных клиентов, которым будет отказано в займе.

F-measure – среднее гармоническое между precision и recall, как правило ближе к наименьшему из них. Также известна как F1-score, F1. Благодаря использованию и точности и полноты при расчете позволяет получить наиболее достоверную оценку точности классификации в целом, следовательно высокое (близкое к 1) значение F1 говорит о правильной классификации всех классов объектов. Считается одной из лучших метрик для оценки качества бинарной классификации, особенно если рассматриваются несбалансированные данные и количество одного класса заметно (в 10 и более раз) превосходит количество другого класса.

В рамках примера с кредитным скорингом F1 является наилучшей метрикой точности, особенно если бизнес-задача постановлена таким образом, что необходимо точно определить ненадежных клиентов, при этом надежные клиенты должны ошибочно определяться как ненадежные как можно реже.

# **2. Обработка исходных данных с помощью логистической регрессии и ансамбля деревьев решений**

В работе рассматривается задача кредитного скоринга, имеются сведения о 2709 клиентах банка, из которых 500 являются ненадежными, цель построение модели для классификации ненадежных клиентов. Для этих целей используются модели логистической регрессии и ансамбль решающих деревьев.

С целью повышения качества моделей, при загрузке исходных данных все столбцы с категориальными переменными заменены столбцами с dummi-переменными.

Первой применяется модель на основе ансамбля деревьев решений, в результате демонстрируется относительно высокая точность распознания надежных клиентов, при низкой (около 50%) точности классификация ненадежных клиентов, результаты можно визуализировать в виде матрицы ошибок, Рисунок 2.

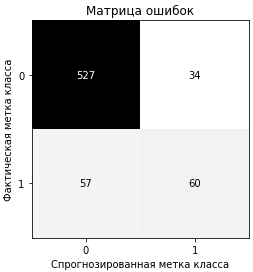


Рисунок 2 Матрица ошибок ансамбля деревьев решений (исходные данные)

Более подробно результат работы ансамбля деревьев решений на исходных данных показывает применение всех ранее рассмотренных метрика качества классификации, таблица 3.

Таблица 3

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.8657817109144543 |
| precision | 0.6382978723404256 |
| recall | 0.5128205128205128 |
| F-measure | 0.5687203791469194 |

Логистическая регрессия демонстрирует несколько худший результат в сравнении с ансамблем деревьев решений, аналогично результаты можно визуализировать в виде матрицы ошибок, Рисунок 2.

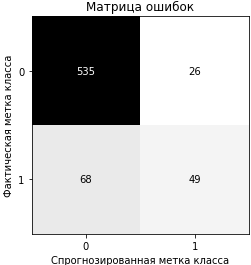


Рисунок 3 Матрица ошибок логистической регрессии (исходные данные)

Более подробно результат работы логистической регрессии на исходных данных показывает применение всех ранее рассмотренных метрика качества классификации, таблица 3.

Таблица 3

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.8613569321533924 |
| precision | 0.6533333333333333 |
| recall | 0.4188034188034188 |
| F-measure | 0.51041 |

В результате матрица ошибок и рассматриваемые метрики (за исключением обманчиво-высокой точности) указывают на низкое качество обеих рассматриваемых моделей, что указывает на необходимость принятия дополнительных шагов для повышения качества моделей, одним из таких вариантов является сэмплинг.

# **3. Применение сэмплинга с целью корректировки баланса классов**

Можно предположить, что низкое качество модели и высоким показатель ошибки второго рода связан с преобладанием отрицательного класса (хороший клиент) над положительным (ненадежный клиент), следовательно стоит использовать “oversampling” – увеличить количество примеров меньшего положительного класса в обучающей (но не тестовой!) выборке за счет дублирования объектов меньшего класса.

В результате количество примеров объектов положительного класса в обучающей выборке было увеличено в 4 раза и составляет 2000 записей, тестовая выборка осталась без изменений.

После процедуры “сэмплинга” ансамблю деревьев решений демонстрирует существенно возросшую точность выявления нежелательных клиентов, при этом количество ошибок первого рода существенно изменилось в худшую сторону, Рисунок 4

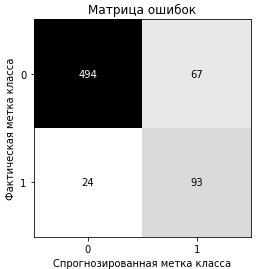


Рисунок 4 Матрица ошибок ансамбля деревьев решений после применения “oversampling”

Более подробно результат работы ансамбля деревьев решений после “oversampling” показывает применение всех ранее рассмотренных метрика качества классификации, таблица 3.

Таблица 3

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.8657817109144543 |
| precision | 0.58125 |
| recall | 0.7948717948717948 |
| F-measure | 0.6714801444043321 |

Схожие результаты демонстрирует и классификатор на базе логистической регрессии, но в отличии от ансамбля деревьев решений, логистическая регрессия показывает несколько меньшую доля ошибок второго рода, при заметно увеличившейся доли ошибок первого рода, Рисунок 5.

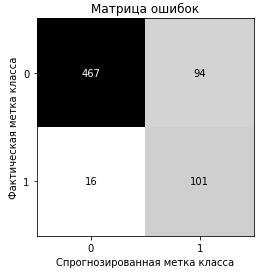


Рисунок 4 Матрица ошибок логистической регрессии после применения “oversampling”

Более подробно результат работы логистической регрессии после “oversampling” показывает применение всех ранее рассмотренных метрика качества классификации, таблица 3.

Таблица 3

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.8377581120943953 |
| precision | 0.517948717948718 |
| recall | 0.8632478632478633 |
| F-measure | 0.6474358974358976 |

# **Выводы**

В результате получены две модели для кредитного скоринга. Модель на базе Логистической регрессии демонстрирует меньший показатель ошибки второго рода и наилучшим образом классифицирует нежелательных клиентов, но это достигается за счет возросшей ошибке первого рода. Соответственно данная модель имеет большее значений полноты (recall) в сравнении с ансамблем решающих деревьев. С другой стороны, ансамблю решающих деревьев хуже классифицирует ненадежных клиентов, но показывает существенно меньшее значение ошибки первого рода – такая модель заметно реже показывает ложноположительный результат, относя надежного клиента к классу ненадежных. Соответственно ансамблю деревьев решений имеет лучшие показатели с точки зрения точности (precision) и F1-меры.

Итоговый выбор модели зависит от бизнес-цели, в случае, когда необходимо минимизировать число ненадежных клиентов, пусть и ценой добросовестных наилучшей выглядит модель на базе логистической регрессии. Но если не менее важно минимизировать случаи отказа выдачи кредита надежному клиенту из-за ложного срабатывания системы, даже если это несколько увеличит риск безвозвратного займа, то можно использовать модель на базе ансамбля решающих деревьев.

При этом необходимо отметить, что точность обеих моделей далека от идеала и при использовании любой из них желательная дополнительная проверка “класса” клиента.

# **Список источников**

1. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — СПб.: Диалектика, 2017. — 480 с.: ил.