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 №7**

**По дисциплине**: Технологии интеллектуального анализа данных

**На тему:** Влияние бизнес-метрик на оценку качества модели бинарной классификации на примере задачи повышения эффективности массовой рассылки клиентам

**Студент: Костромин Д.А.**

**Группа: МИВТ 18-5-7**

**Преподаватель: Фомичева О.Е.**

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

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

Москва, 2019

**Оглавление**

[**Введение** 3](#_Toc25352353)

[**1. Постановка задачи** 4](#_Toc25352354)

[**2. Обучение на исходных данных в условиях преобладания отрицательного класса** 6](#_Toc25352355)

[**3. Применение Over Sampling с целью корректировки баланса классов** 9](#_Toc25352356)

[**Выводы** 11](#_Toc25352357)

[**Список источников** 12](#_Toc25352358)

# **Введение**

В работе рассматриваются особенности оценивания качества моделей в зависимости от конечной цели построения модели. В многих случаях при выборе способов оценивания моделей машинного обучения важную роль играют высокоуровневые цели проекта, в рамках которого будет использована ML-модель, которые также называют бизнес-метриками. В качестве примера используется задача повышения эффективности массовой рассылки клиентам, в рамках которой наиболее важна точная классификация клиентов, которые скорее всего совершат покупку в результате рассылки.

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

# **1. Постановка задачи**

Задача относится к бинарной классификации, в соответствии с которой “положительным” классом является клиент, который совершит покупку после получений информационной рассылки, соответственно к “отрицательному” классу относятся клиенты, которые проигнорируют рассылку. Следовательно, по аналогии с задачей кредитного скоринга одним из способов оценки качества модели послужит матрица ошибок, Рисунок 1. [1]

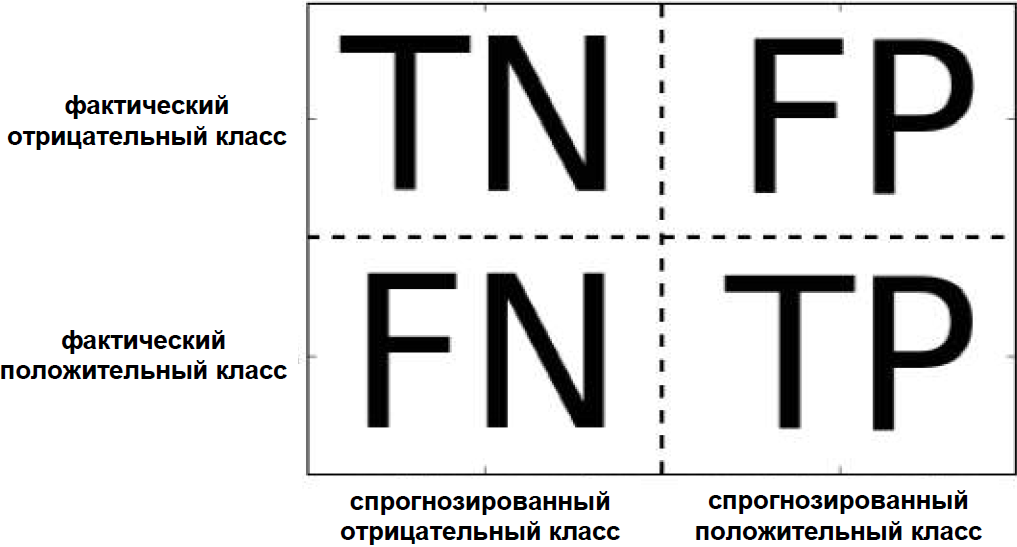


Рисунок 1 Типы ошибок при бинарной классификации

При этом наиболее важным критерием будет выступать ожидаемый доход по результатам рассылки, вычисляемый следующим образом:

Где:

CM = 1.0 – расходы на одну рассылку.

CR = 9.0 – издержки на обслуживание клиента.

R = 20.0 – ожидаемая выручка с одного заказа.

Следовательно, наилучшей метрикой в данной ситуации будет полнота (Recall) , являющийся отношением количества истинно положительных (TP) объектов к общему числу объектов, классифицированных как положительные, включая ошибку второго рода – объектов, ошибочно определенных как положительные (TP + FN). Так как бизнес-метрики указывают на необходимость минимизации ошибки второго рода (FN) и максимально – точной классификации истинно-положительного класса (TP), или этом величина ошибки первого рода (FP) в соответствии с бизнес метриками является гораздо менее важной, чем ошибка второго рода, так как предприятие несет существенно меньшие убытки в “ложной” случае отправки информационной рассылки, чем если рассылка не будет отправлена готовому к покупки клиенту. Но это не означает что можно полностью проигнорировать ошибку первого рода.

Исходные данные содержатся в двух файлах, responses\_train.txt и responses\_test.txt, представляющих из себя обучающую и отложенную выборку соответственно. При выполнении работы будут рассмотрены модели дерева решений, случайного леса и логистической регрессии, которые будут обучены и протестированы на данных из файла responses\_train.txt, а затем настроенные модели будут использованы для дополнительного тестирования на отложенной выборке.

# **2. Обучение на исходных данных в условиях преобладания отрицательного класса**

Первоначально осуществляется обучение моделей на исходных данных, в которых доля примеров положительного класса составляет 14,5%.

Матрица ошибок дерева решений демонстрирует высокую точность классификации отрицательного класса, при этом заметен высокий показатель ошибки второго рода – в результате, при использовании такой модели рассылка не будет отправлена “лишним” клиентам, но и многие из готовых к покупке клиентов будут проигнорированы, что недопустимо, Рисунок 1.

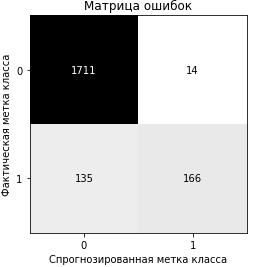


Рисунок 2 Матрица ошибок дерева решений (исходные данные)

Более подробно результат работы дерева решений на исходных данных показывает применение всех ранее рассмотренных метрика качества классификации и ожидаемого дохода, Таблица 1.

Таблица 1 Метрики оценки дерева решений (исходные данные)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.9264560710760118 |
| precision | 0.9222222222222223 |
| recall | 0.5514950166112956 |
| F-measure | 0.6902286902286902 |
| Доход | 1646 |

Схожие результаты демонстрирует случайные лес, единственными отличиями от дерева решений являются незначительно увеличившаяся ошибка первого рода, и незначительно снизившаяся ошибка второго рода, Рисунок 2.

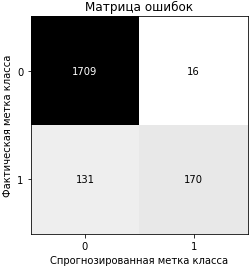


Рисунок 2 Матрица ошибок случайного леса (исходные данные)

Более подробно результат работы случайного леса на исходных данных показывает применение всех ранее рассмотренных метрика качества классификации и ожидаемого дохода. Случайный лес демонстрирует увеличение дохода в сравнении с деревом решений, но разница настолько незначительна что ей можно пренебречь, Таблица 2.

Таблица 2 Метрики оценки случайного леса (исходные данные)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.9274432379072063 |
| precision | 0.9139784946236559 |
| recall | 0.5647840531561462 |
| F-measure | 0.6981519507186857 |
| Доход | 1684 |

Логистическая регрессия демонстрирует заметно худший результат с незначительно возросшей долей ошибки первого рода и существенно увеличившимся показателем ошибки второго рода, Рисунок 3.

Более подробно результат работы логистической регрессии на исходных данных показывает применение всех ранее рассмотренных метрика качества классификации и ожидаемого дохода. Как и следовало ожидать, логистическая регрессия показывает значительно более низкий уровень прогнозируемого дохода в сравнении с деревом решений и случайным лесом, Таблица 3.

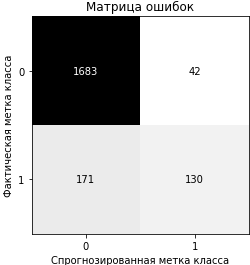


Рисунок 3 Матрица ошибок логистической регрессии (исходные данные)

Таблица 3 Метрики оценки случайного леса (исходные данные)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.8948667324777887 |
| precision | 0.7558139534883721 |
| recall | 0.4318936877076412 |
| F-measure | 0.5496828752642706 |
| Доход | 1258 |

# **3. Применение Over Sampling с целью корректировки баланса классов**

Можно предположить, что высокий уровень точности классификации отрицательно класса при условии низкой точности классификации положительного класса связан с несбалансированностью исходных данных и, по аналогии с задачей кредитного скоринга, одним из возможных способов разрешения данной проблемы является Over Sampling – увеличение доли меньшего класса в обучающей выборке с помощью дублирования примеров меньшего класса.

При увеличении доли меньшего класса в обучающей выборке демонстрируется уменьшение количества ошибок второго рода для всех моделей, ценой роста ошибки первого рода. Наилучшие результаты демонстрируются при увеличении доли меньшего класса в обучающей выборке в 7 раз.

По аналогии с предыдущим этапом, первым демонстрируются оценки для модели на базе дерева решений. В результате демонстрируется снижение ошибки второго рода ценой возрастания ошибки первого рода, Рисунок 4

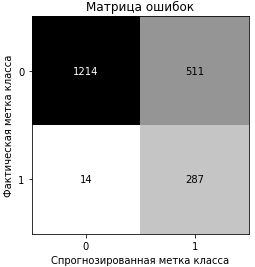


Рисунок 4 Матрица ошибок дерева решений (Over Sampling)

Более подробно результат работы дерева решений после Over Sampling показывает применение всех ранее рассмотренных метрика качества классификации и ожидаемого дохода. В результате не смотря на заметно возросший процент ошибки первого рода, демонстрируется рост ожидаемого дохода, так как с точки зрения бизнес-метрик снижение доли ошибки второго рода несколько важнее увеличения доли ошибки первого рода, пока она сохранятся в разумных пределах, также демонстрируется высокий показатель метрики recall, наиболее подходящей для оценки модели в соответствии с имеющимися бизнес-метриками, Таблица 4.

Таблица 4 Метрики оценки дерева решений (Over Sampling

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.7408687068114511 |
| precision | 0.35964912280701755 |
| recall | 0.9534883720930233 |
| F-measure | 0.5222929936305734 |
| Доход | 2359 |

Модель на базе случайного леса, как и на этапе до Over Sampling, практически полностью соответствует дереву решений, за исключением незначительно уменьшения доли FN-ошибки и незначительного увеличения FP-ошибки, Рисунок 5.

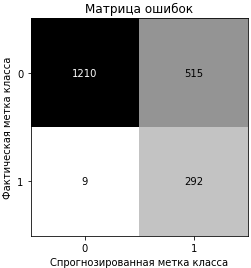


Рисунок 5 Матрица ошибок случайного леса (Over Sampling)

Несколько больше о отличиях между результатами случайного леса и дерева решения могут сказать метрики. Случайные лес показывает большие показателя метрики recall и прогнозируемого дохода, но отличия от дерева решений крайне незначительны, Таблица 5

Таблица 5 Метрики оценки случайного леса (Over Sampling)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.7413622902270484 |
| precision | 0.3618339529120198 |
| recall | 0.9700996677740864 |
| F-measure | 0.5222929936305734 |
| Доход | 2405 |

Аналогично предыдущему этапу наихудший результат показала логистическая регрессия, данная модель демонстрирует несколько меньший показатель ошибки первого рода в сравнении с другими моделями, но при этом демонстрирует значительно большую ошибку второго рода, Рисунок 6.

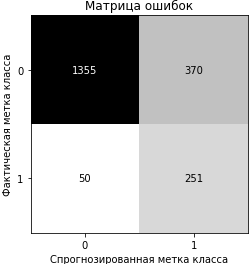


Рисунок 6 Матрица ошибок логистической регрессии (исходные данные)

Метрики качества логистической регрессии демонстрируют несколько большие показатели F1, accuracy и precision, менее важных в данной задачи, при этом логистическая регрессия демонстрирует заметно худшие показатели прогнозируемого дохода и precision, наиболее важных в данном случае, Таблица 6.

Таблица 6 Метрики оценки логистической регрессии (Over Sampling)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.792694965449161 |
| precision | 0.40418679549114334 |
| recall | 0.8338870431893688 |
| F-measure | 0.544468546637744 |
| Доход | 2140 |

# **4. Тестирование моделей на отложенной выборке**

Теперь, когда имеются обученные модели можно проверить их обобщающую способность на неиспользованных ранее данных из файла responses\_test.txt.

Аналогично первым применяется дерево решений, обобщающая способность которых на отложенной выборке подтверждает результат, полученный на тестовой выборке, Рисунок 7.

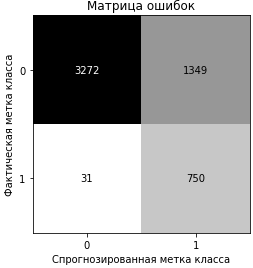


Рисунок 7 Матрица ошибок дерева решений (отложенная выборка)

Метрика качества дерева решений уточняют оценку обобщающей способности, но результаты практически идентичны тестовой выборке, Таблица 7

Таблица 7 Метрики оценки дерева решений (отложенная выборка)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.7445390596075527 |
| precision | 0.35731300619342543 |
| recall | 0.9603072983354674 |
| F-measure | 0.5208333333333334 |
| Доход | 6151 |

Как и на тестовой выборке, результаты случайного леса при проверке на отложенной выборке незначительно лучше результатов дерева решений, Рисунок 8.

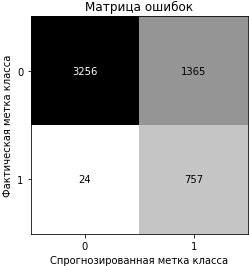


Рисунок 8 Матрица ошибок случайного леса (отложенная выборка)

Метрики качества случайного леса на отложенной выборке, по аналогии с деревом решений, слабо отличаются от своих аналогов на тестовой выборке, Таблица 8.

Таблица 8 Метрики оценки случайного леса (отложенная выборка)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.7428730099962977 |
| precision | 0.3567389255419416 |
| recall | 0.969270166453265 |
| F-measure | 0.5215294522907338 |
| Доход | 6205 |

Результаты логистической регрессии также соответствуют результатам на тестовой выборке с сохранением имевшихся недостатков, Рисунок 9

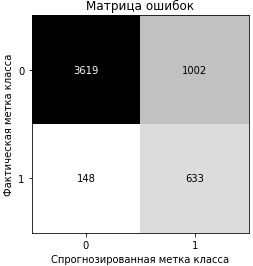


Рисунок 8 Матрица ошибок логистической регрессии (отложенная выборка)

Метрики оценки качества уточняют характеристики модели логистической регрессии после тестирования на отложенной выборке, но в целом результат неотличим от проверки на тестовой выборке, Таблица 9

Таблица 9 Метрики оценки логистической регрессии (отложенная выборка)

|  |  |
| --- | --- |
| Метрика | Оценка |
| accuracy | 0.787115883006294 |
| precision | 0.3871559633027523 |
| recall | 0.8104993597951344 |
| F-measure | 0.5240066225165563 |
| Доход | 5328 |

В результате модели дерево и случайный лес демонстрируют наилучший и практически идентичный результат, тогда как классификация клиентов с помощью логистической регрессии показывает заметно худшие результаты, измеряемые как прогнозируемый доход, Рисунок 10.

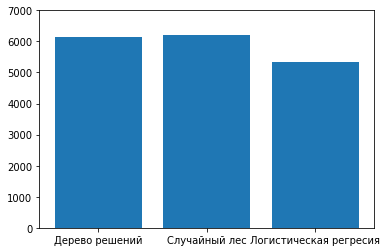


Рисунок 10 Сопоставление результатов по бизнес-метрике прогнозируемого дохода

# **Выводы**

В результате получены модели классификации клиентов на базе случайного леса и дерева решений, показавшие наилучшие результаты по бизнес-метрике прогнозируемого дохода, тогда так логистическая регрессия показала наихудший результат. При этом необходимо учесть несколько факторов, в первую очередь все полученные модели имеют далекую от идеальной обобщающую способность, о чем говорит высокий показатель ошибки первого рода и низкие оценки метрики F1-среднее, но случайный лес и дерево решений имеют высокий показатель полноты (recall), указывающий на низкую долю ошибки второго рода, важную с точки зрения бизнес правил, что и делает их, не смотря на достаточно низкий показатель обобщенной точности подходящими для решения данной задачи в соответствии с бизнес-метриками, указывающими что ложно-положительное срабатывание модели имеет относительно небольшие негативные последствия, в сравнении с ложно-отрицательной классификацией, приводящей к заметным убыткам из за потери потенциальных клиентов.

В результате наилучшей моделью для классификации клиентов с целью оптимизация рассылки является случайный лес.

# **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.