ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА (САМАРСКИЙ УНИВЕРСИТЕТ)»

ИНСТИТУТ ИНФОРМАТИКИ И КИБЕРНЕТИКИ
Кафедра программных систем

Дисциплина
Системы искусственного интеллекта

ОТЧЕТ

по лабораторной работе № 2

Задача классификации

Выполнили: Фокин Е.А., Сидоров А.О., Павлов В.О., Мергалиев Р.Е., Хасанов Д.И., Клеймёнов А. С., Сальников И.А., группа № 6303-030203D

Проверила: Жданова А.Н.

Самара 2025

ВВЕДЕНИЕ

Целью данной лабораторной работы являлось построение и всесторонняя оценка моделей бинарной классификации для решения задачи прогнозирования согласия клиента на открытие банковского вклада. В ходе работы необходимо было освоить и применить базовые и продвинутые алгоритмы машинного обучения, метрики качества для несбалансированных выборок, а также методы интерпретации и анализа ошибок моделей.

1. Подготовка и предобработка данных

Исходный датасет содержал информацию о клиентах банка, параметрах маркетинговой кампании и макроэкономических показателях. Перед построением моделей была выполнена комплексная предобработка данных.

Этапы предобработки:

1. Удаление дубликатов: Из набора данных было удалено 12 полностью дублирующихся записей, что обеспечило уникальность каждого наблюдения;
2. Обработка скрытых пропусков: Значения 'unknown', представляющие собой скрытые пропуски, были заменены на самое частое значение (моду) в соответствующем категориальном столбце;
3. Преобразование типов данных: Все столбцы типа object были преобразованы в категориальный тип для оптимизации использования памяти (сокращение с 6.6+ MB до 3.9 MB) и семантической корректности;
4. Преобразование целевой переменной: Целевая переменная y была преобразована из категориального формата (yes/no) в числовой (1/0) для корректной работы алгоритмов машинного обучения.

```
Первые 5 строк обработанного датасета:
============================================================
   age        job  marital    education default housing loan    contact month  \
0   56  housemaid  married     basic.4y      no      no   no  telephone   may
1   57   services  married  high.school      no      no   no  telephone   may
2   37   services  married  high.school      no     yes   no  telephone   may
3   40     admin.  married     basic.6y      no      no   no  telephone   may
4   56   services  married  high.school      no      no  yes  telephone   may

  day_of_week  ...  campaign  pdays  previous     poutcome emp.var.rate  \
0         mon  ...         1    999         0  nonexistent          1.1
1         mon  ...         1    999         0  nonexistent          1.1
2         mon  ...         1    999         0  nonexistent          1.1
3         mon  ...         1    999         0  nonexistent          1.1
4         mon  ...         1    999         0  nonexistent          1.1

   cons.price.idx  cons.conf.idx  euribor3m  nr.employed  y
0          93.994          -36.4      4.857       5191.0  0
1          93.994          -36.4      4.857       5191.0  0
2          93.994          -36.4      4.857       5191.0  0
3          93.994          -36.4      4.857       5191.0  0
4          93.994          -36.4      4.857       5191.0  0

[5 rows x 21 columns]
============================================================

Итоговая размерность: (41176, 21)
```

Рисунок 1 - Пример очищенных данных после предобработки

Обоснование выбора стратегии замены на моду:

Для категориальных данных замена на моду является оптимальной стратегией по следующим причинам:

1. По сравнению с медианой: Медиана неприменима к категориальным данным, так как они не имеют числового порядка;
2. По сравнению с константой: Замена на специальное значение создает искусственную категорию, которая может ввести модель в заблуждение;
3. Преимущества моды: Сохраняет исходное распределение категорий, не создает артефактов, обеспечивает минимальное искажение исходных закономерностей;

```
Анализ и обработка 'unknown'
Значения 'unknown' являются скрытыми пропусками. Проанализируем их количество.
- Столбец 'job': 330 значений 'unknown' (0.80%)
- Столбец 'marital': 80 значений 'unknown' (0.19%)
- Столбец 'education': 1730 значений 'unknown' (4.20%)
- Столбец 'default': 8596 значений 'unknown' (20.88%)
- Столбец 'housing': 990 значений 'unknown' (2.40%)
- Столбец 'loan': 990 значений 'unknown' (2.40%)
```

Рисунок 2 - Распределение пропущенных значений ('unknown') по столбцам датасета

Для обеспечения корректной и воспроизводимой обработки признаков был создан единый конвейер (Pipeline) с помощью ColumnTransformer, который выполнял:

1. Масштабирование числовых признаков методом StandardScaler;
2. Кодирование категориальных признаков методом OneHotEncoder.

Это позволило инкапсулировать всю логику предобработки и избежать утечки данных при обучении и валидации моделей.

2. Разбиение данных и стратегия валидации

Для объективной оценки моделей данные были разделены на три выборки в соотношении 60/20/20 (обучающая, валидационная и тестовая соответственно). Было применено стратифицированное разбиение по целевой переменной y для сохранения исходной пропорции классов (~11.3% положительного класса) во всех выборках, что критически важно в условиях дисбаланса. Все случайные процессы были зафиксированы с помощью random_state=42 для обеспечения полной воспроизводимости экспериментов.

3. Построение и сравнение моделей

На валидационной выборке было проведено сравнение нескольких моделей. В качестве ключевых метрик использовались F1-score (для класса 1), ROC-AUC и PR-AUC, так как они наиболее адекватно отражают качество классификации на несбалансированных данных.

3.1. Базовые модели

1. Logistic Regression: Простая, но надежная линейная модель. Показала F1-score = 0.58, ROC-AUC = 0.9359, PR-AUC = 0.5873;
2. Decision Tree: Для дерева решений с помощью GridSearchCV была подобрана оптимальная глубина (max_depth=10). Модель показала F1-score = 0.56, ROC-AUC = 0.8877, PR-AUC = 0.5405.

Логистическая регрессия была выбрана в качестве baseline, так как продемонстрировала лучшие результаты по всем ключевым метрикам.

4. Продвинутая модель

В качестве продвинутой модели был выбран LightGBM — эффективная реализация градиентного бустинга. Для подбора оптимальных гиперпараметров использовался RandomizedSearchCV, который протестировал 10 случайных конфигураций.

LightGBM (с настроенными параметрами): Модель показала F1-score = 0.63, ROC-AUC = 0.9445, PR-AUC = 0.6550.

Настроенная модель LightGBM продемонстрировала существенное превосходство над baseline по всем метрикам и была выбрана в качестве финальной ("победителя").

```
Отчет по Логистической регрессии на валидационной выборке:
              precision    recall  f1-score   support

           0       0.98      0.86      0.91      7307
           1       0.44      0.87      0.58       928

    accuracy                           0.86      8235
   macro avg       0.71      0.87      0.75      8235
weighted avg       0.92      0.86      0.88      8235

ROC-AUC: 0.9359
PR-AUC: 0.5873
```

```
Отчет по LightGBM (с лучшими параметрами) на валидационной выборке:
              precision    recall  f1-score   support

           0       0.98      0.89      0.93      7307
           1       0.50      0.84      0.63       928

    accuracy                           0.89      8235
   macro avg       0.74      0.87      0.78      8235
weighted avg       0.92      0.89      0.90      8235

ROC-AUC: 0.9445
PR-AUC: 0.6550
```

Рисунок 3, 4 – Сравнение метрик качества для baseline-модели (Logistic Regression) и продвинутой модели (LightGBM) на валидационной выборке.

5. Финальная оценка и интерпретация модели-победителя

Лучшая модель (LightGBM) была оценена на отложенной тестовой выборке для получения несмещенной оценки ее производительности.

5.1. Результаты на тестовой выборке

1. F1-score (class 1): 0.64;
2. ROC-AUC: 0.9470;
3. PR-AUC: 0.6454.

```
Финальный отчет по LightGBM на тестовой выборке:
              precision    recall  f1-score   support

           0       0.98      0.89      0.93      7308
           1       0.51      0.86      0.64       928

    accuracy                           0.89      8236
   macro avg       0.74      0.88      0.79      8236
weighted avg       0.93      0.89      0.90      8236

ROC-AUC: 0.9470
PR-AUC: 0.6454
```

Рисунок 5 – Финальные метрики качества модели LightGBM на отложенной тестовой выборке.

Результаты на тестовых данных оказались очень близки к валидационным, что свидетельствует об отсутствии переобучения и хорошей обобщающей способности модели.

5.2. Интерпретация модели (Важность признаков)

Для интерпретации предсказаний был использован метод Permutation Importance. Анализ показал, что наибольший вклад в качество модели вносят следующие признаки:

1. **duration**: Длительность последнего телефонного контакта;
2. **euribor3m**: Межбанковская ставка Euribor, отражающая экономическую ситуацию;
3. **emp.var.rate**: Квартальный уровень изменения занятости.

	feature	importance_mean
1	num__duration	0.087470
8	num__euribor3m	0.055767
5	num__emp.var.rate	0.043286
9	num__nr.employed	0.006824
0	num__age	0.001736
44	cat__day_of_week_wed	0.001178
42	cat__day_of_week_thu	0.001044
24	cat__education_high.school	0.000741
30	cat__loan_yes	0.000656
29	cat__housing_yes	0.000546

Рисунок 6 – Важность признаков, рассчитанная методом Permutation Importance. Наибольшее влияние на предсказание оказывают duration, euribor3m и emp.var.rate.

Это подтверждает гипотезу о том, что поведенческие и макроэконо1мические факторы являются ключевыми предикторами для данной задачи.

5.3. Анализ ошибок

Найдено 781 ложноположительных ошибок (FP). Примеры:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	previous	poutcome
35994	31	blue-collar	single	professional.course	no	no	no	cellular	may	tue	...	0	nonexistent
27173	40	blue-collar	divorced	basic.9y	no	no	no	cellular	nov	fri	...	0	nonexistent
24421	52	admin.	divorced	high.school	no	no	no	cellular	nov	mon	...	0	nonexistent

3 rows × 23 columns

Найдено 128 ложноотрицательных ошибок (FN). Примеры:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	previous	poutco
27704	44	blue-collar	single	professional.course	no	no	no	cellular	mar	mon	...	0	nonexist
38483	55	unemployed	married	basic.9y	no	no	no	cellular	oct	tue	...	0	nonexist
30027	64	retired	married	basic.4y	no	yes	no	cellular	apr	wed	...	2	fail

3 rows × 23 columns

Рисунок 7 – Примеры ложноотрицательных срабатываний (FN). Вероятность для первого примера (0.488) близка к порогу классификации 0.5.

Анализ ложноположительных (FP) и ложноотрицательных (FN) срабатываний показал:

1. FP (ложная тревога): Модель склонна ошибочно предсказывать согласие для клиентов с аномально долгими, но безуспешными разговорами;
2. FN (пропуск цели): Самая "дорогая" ошибка. Модель часто упускает клиентов, которые согласились, но не соответствовали типичному "успешному" профилю (например, короткий звонок). Анализ показал, что часть таких ошибок происходит при вероятности, близкой к порогу 0.5.

6. Итоговые выводы

1. В ходе лабораторной работы была успешно построена и оценена модель градиентного бустинга (LightGBM), способная с высокой точностью (PR-AUC = 0.6454) прогнозировать отклик клиентов на маркетинговое предложение банка;
2. Доказано, что продвинутая модель LightGBM значительно превосходит базовые подходы (Логистическая регрессия, Дерево решений) для решения данной задачи;
3. Интерпретация модели показала, что ключевыми факторами для прогнозирования являются динамические признаки (длительность контакта, экономические индикаторы), а не статичные демографические данные;
4. Анализ ошибок позволил сформулировать конкретные шаги для дальнейшего улучшения модели: создание новых, более сложных признаков (Feature Engineering) и подбор оптимального порога классификации для минимизации количества "пропущенных" клиентов;
5. Вся работа была выполнена с соблюдением принципов воспроизводимости, включая фиксацию random_state и версий библиотек;