Министерство образования и науки РФ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Омский государственный технический университет»

Кафедра «Автоматизированные системы обработки информации и управления»

Курсовой проект по дисциплине

«Динамические языки программирования»

Выполнил

Студент гр. ПИН-202 Духовский В.О.

(подп., дата)

Проверил

Старший преподаватель каф. АСОИУ Кабанов А.А.

(подп., дата)

Омск 2023

**Реферат**

ОТЧЕТ 31 с., 13 рис., 3 источника.

КУРСОВОЙ ПРОЕКТ, МАШИННОЕ ОБУЧЕНИЕ, PYTHON, CATBOOST

Цель курсовой работы — ознакомиться и приобрести базовые знания в области машинного обучения.

В данной работе выполнены:

1. метод к- ближайших соседей;
2. метод машины опорных векторов;
3. методы линеной и логистической регрессий;
4. метод наивного Байеса;
5. методы решающего дерева и случайного леса;
6. метод CatBoost.

СОДЕРЖАНИЕ

[1. Метод к-ближайших соседей (k-NN) 6](#_Toc154705613)

[1.1 Описание метода 6](#_Toc154705614)

[1.2 Принцип работы 6](#_Toc154705615)

[1.3 Применение на наборе данных 6](#_Toc154705616)

[1.4 Преимущества и ограничения 10](#_Toc154705617)

[2. Метод машины опорных векторов (SVM) 11](#_Toc154705618)

[2.1 Описание метода 11](#_Toc154705619)

[2.2 Разделительная гиперплоскость и принцип работы 11](#_Toc154705620)

[2.3 Применение на наборе данных 11](#_Toc154705621)

[2.4 Преимущества и ограничения 12](#_Toc154705622)

[3. Методы линейной и логистической регрессии 15](#_Toc154705623)

[3.1 Описание методов 15](#_Toc154705624)

[3.2 Принцип работы 16](#_Toc154705625)

[3.3 Сравнительный анализ с другими методами 18](#_Toc154705626)

[4. Метод наивного Байеса 19](#_Toc154705627)

[4.1 Описание метода 19](#_Toc154705628)

[4.2 Применение на наборе данных 19](#_Toc154705629)

[4.3 Преимущества и ограничения 21](#_Toc154705630)

[5.1 Описание метода 22](#_Toc154705631)

[5.2 Применение на модельных данных 23](#_Toc154705632)

[5.3 Сравнение с другими методами классификации 25](#_Toc154705633)

[6. Метод CatBoost 26](#_Toc154705634)

[6.1 Описание метода 26](#_Toc154705635)

[6.2 Применение на модельных данных 27](#_Toc154705636)

[Заключение 29](#_Toc154705637)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 31](#_Toc154705638)

**Введение**

Машинное обучение — одна из наиболее динамично развивающихся областей, которая позволяет компьютерам обучаться на основе данных и делать прогнозы или принимать решения без явного программирования.

Цель настоящей курсовой работы заключается в ознакомлении с основами машинного обучения и рассмотрении различных методов классификации. В ходе исследования были проведены эксперименты с различными алгоритмами, такими как метод к-ближайших соседей, машина опорных векторов, линейная и логистическая регрессии, наивный Байес, а также решающее дерево и случайный лес.

В рамках данной работы представлен обзор и сравнение указанных методов классификации на простых модельных данных. Каждый из методов исследован с целью понимания их принципов работы, преимуществ и ограничений.

# Метод к-ближайших соседей (k-NN)

# Описание метода

Метод к-ближайших соседей (k-NN) относится к одному из простейших алгоритмов классификации в машинном обучении. Он основывается на принципе близости объектов: если у объекта есть соседи известного класса, то скорее всего, этот объект также принадлежит к этому классу.

# Принцип работы

При классификации нового объекта метод k-NN ищет k ближайших к нему объектов в обучающем наборе данных. Затем присваивает новому объекту тот класс, который наиболее часто встречается среди его соседей.

# Применение на наборе данных

Для обучения моделей был выбран датасет. Датасет представляет собой коллекцию данных, предназначенных, для анализа кредитных заявок или риска мошенничества в банковской сфере. Описание полей данных:

1. title (заголовок): Описание компаний.
2. company (Компания): Название компаний.
3. salary (Зарплата): Размер заработной платы в компаниях.
4. rating (рейтинг): Рейтинг компаний.
5. review\_count (количество просмотров):
6. types (Типы): Тип компании.
7. location (местоположение): Местоположение компании.
8. hires\_needed (требуются сотрудники): Требуются ли человек в данную компанию.
9. hires\_needed\_exact (количество сотрудников): Количество человек необходимых компаниям.
10. urgently\_hiring (срочно требуются сотрудники): Как срочно требуются люди в данную компанию.
11. remote\_work\_model (удалённая работа): Работа происходит в удалённом формате .
12. snippet (описание найма): Описание вакансии .
13. job\_link (ссылка на вакансию): Ссылка на вакансию.
14. sponsored (спонсируемая): Является ли компания спонсируемой.
15. featured\_employer (рекомендуемый работодатель): Отзывы о работодателе и о самой компании.
16. ad\_id (идентификатор объявления): ID объявления
17. source\_id (идентификатор источника):ID работодателя
18. job\_location\_postal (почтовый адрес места работы): Адрес электронной почты компании.
19. company\_overview\_link (ссылка на сайт компании): Ссылка на сайт компании.
20. location\_extras (дополнительные сведения о местоположении)
21. hiring\_event\_job (Целевой столбец): Является ли данная компания подходящей для человека.

На рисунке 1-2 представлены подбор гиперпараметров и наилучшего k

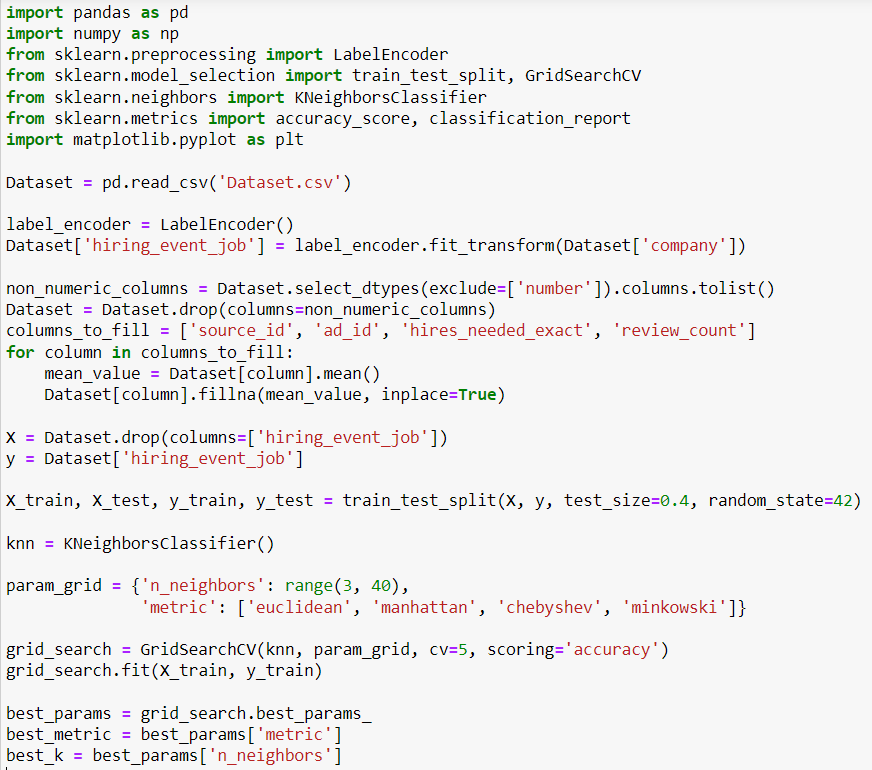


Рис. 1 – Подбор гиперпараметра и наилучшего k

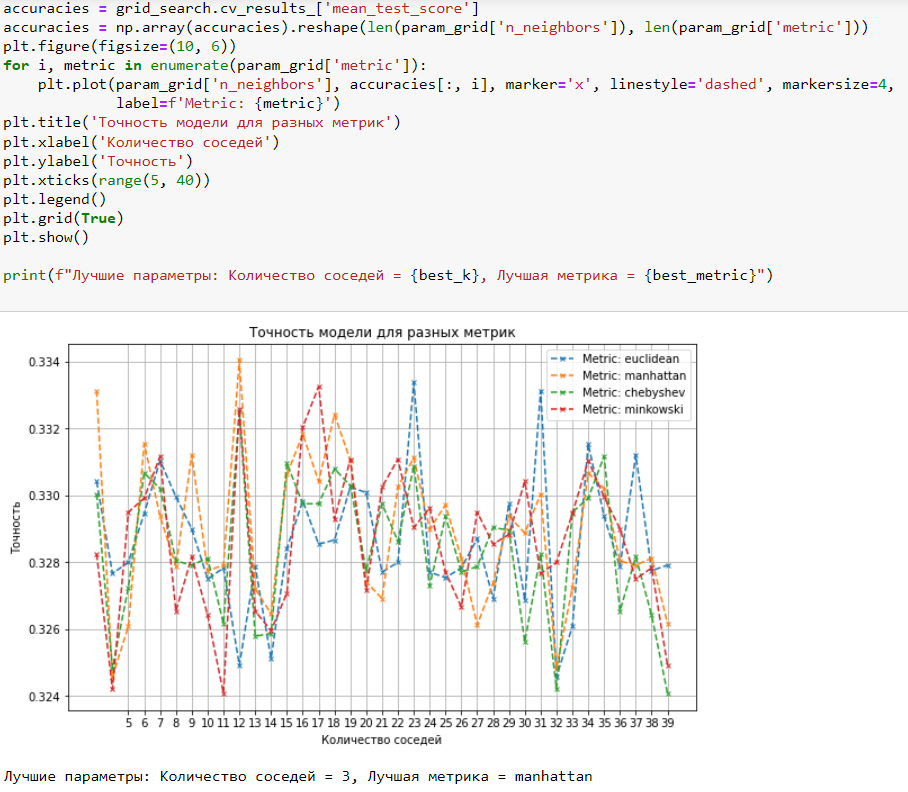


Рис. 2 – Подбор гиперпараметра и наилучшего k

На рисунке 3 представлены проверка на тестовых данных и вывод результатов

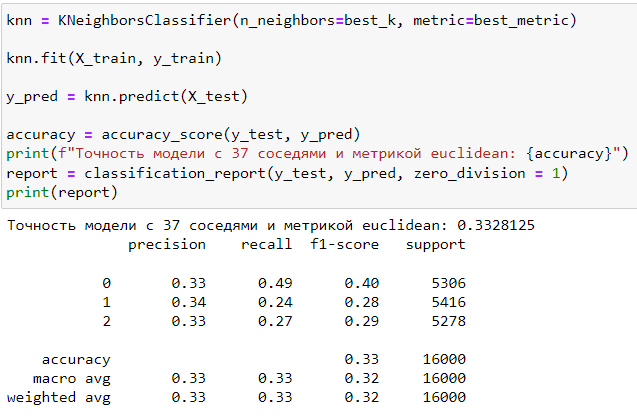


Рис.3 – Проверка на тестовых данных и вывод результатов

# Преимущества и ограничения

Преимущества метода k-NN:

* Прост в реализации и понимании.
* Не требует обучения модели, пока поступают новые данные.
* Хорошо подходит для начальной оценки данных и быстрых прототипов.

Однако у метода k-NN есть и ограничения:

* Чувствителен к выбросам в данных.
* Неэффективен на больших наборах данных из-за вычислительной сложности.
* Не учитывает значимость признаков, все признаки равнозначны.

Метод k-NN полезен для первичного анализа данных, но в реальных приложениях может потребоваться более сложная модель для учета различных аспектов и повышения точности предсказаний.

# Метод машины опорных векторов (SVM)

# Описание метода

Метод машины опорных векторов (SVM) является мощным алгоритмом машинного обучения, используемым для задач классификации и регрессии. Основная идея заключается в поиске оптимальной разделительной гиперплоскости, которая максимально разделяет классы в данных.

# Разделительная гиперплоскость и принцип работы

SVM строит гиперплоскость в n-мерном пространстве, где n - количество признаков. Эта гиперплоскость разделяет пространство на две части и максимизирует расстояние (зазор) между объектами разных классов, называемое отступом. Оптимальная гиперплоскость выбирается так, чтобы этот отступ был максимальным.

# Применение на наборе данных

На рисунке 4 представлено описание гиперпараметров и подбор гиперпараметров с помощью перекрестной проверки

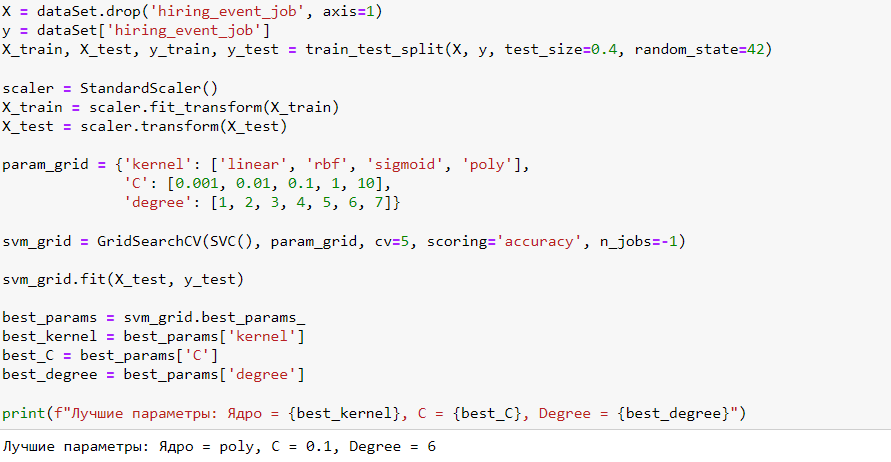


Рис.4 – Подбор гиперпараметров с помощью перекрестной проверки

На рисунке 5 представлена оценка производительности модели на тестовом наборе

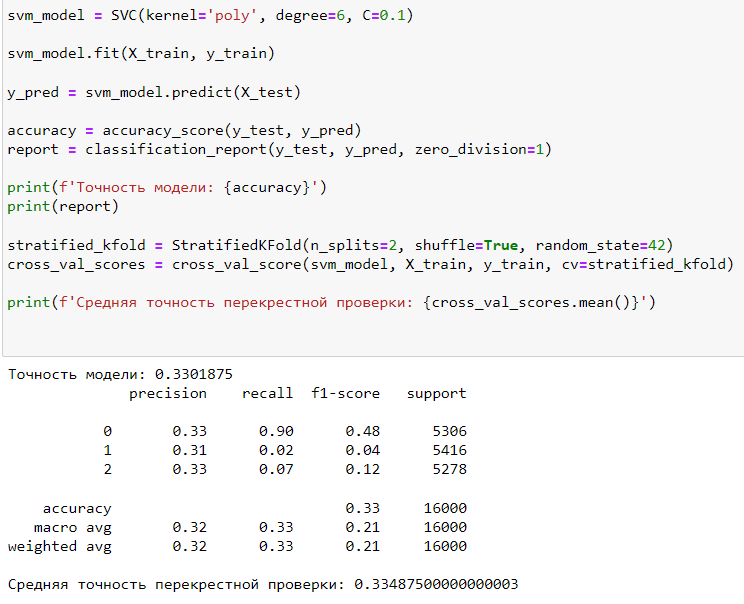


Рис. 5 - Оценка производительности модели на тестовом наборе

# Преимущества и ограничения

Преимущества метода SVM:

* Эффективен в пространствах с большим количеством признаков.
* Хорошо работает в условиях разделяемых данных с помощью нелинейных ядер.
* Стабилен при обучении на небольшом наборе данных.

Однако у метода SVM есть и ограничения:

* Требует тщательного выбора гиперпараметров и ядра.
* Неэффективен при работе с большими наборами данных из-за вычислительной сложности.
* Может быть чувствителен к выбросам в данных.

SVM - мощный алгоритм, который может быть эффективен при правильной настройке, но требует опыта для правильного применения и настройки гиперпараметров для достижения оптимальных результатов.

# Методы линейной и логистической регрессии

# Описание методов

Линейная регрессия – это метод, используемый для прогнозирования значений непрерывной зависимой переменной на основе линейной комбинации независимых переменных. Основная идея заключается в поиске линейной зависимости между предикторами и целевой переменной.

Применение линейной регрессии может быть полезным при прогнозировании численных значений, например, прогнозировании цены на недвижимость на основе её характеристик, таких как площадь, количество комнат и т.д.

Логистическая регрессия используется для задач классификации, прогнозируя вероятность принадлежности объекта к определенному классу. В отличие от линейной регрессии, логистическая регрессия применяется для бинарной или многоклассовой классификации.

# Принцип работы

На рисунке 6 представленно обучение модели модели при помощи линейной регрессии

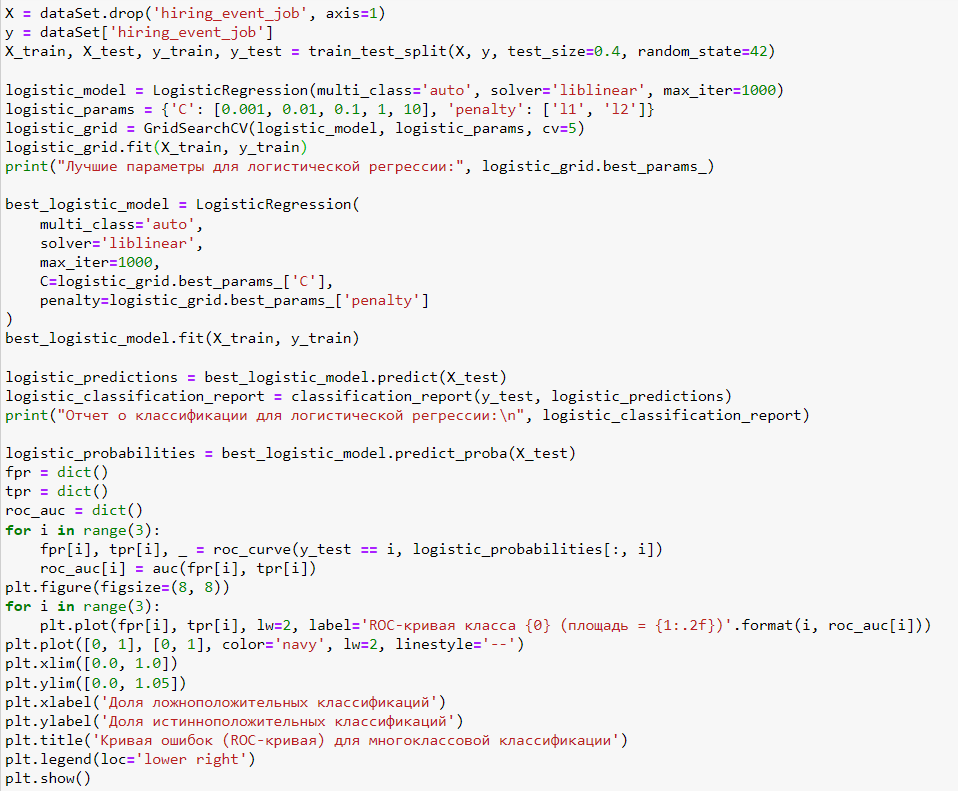


Рис.6 – Обучение модели методом линейной регрессии

На рисунке 7 представлен вывод результатов

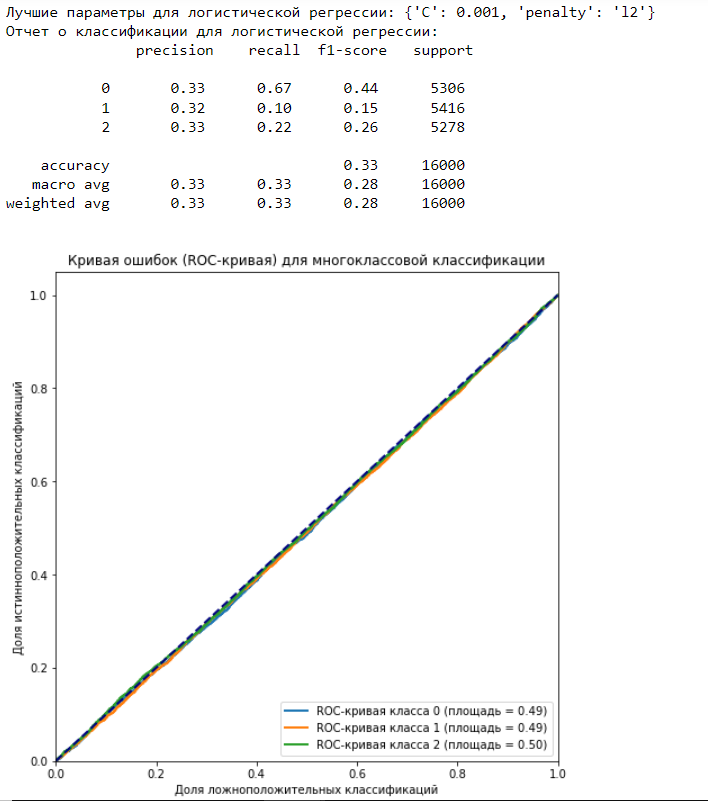


Рис.7 – Вывод результатов

# Сравнительный анализ с другими методами

Линейная и логистическая регрессии имеют свои сильные и слабые стороны по сравнению с другими методами. Линейная регрессия хорошо работает при предсказании непрерывных значений, тогда как логистическая регрессия применяется для задач классификации. Однако оба метода чувствительны к выбросам в данных и могут быть недостаточно гибкими для улавливания сложных нелинейных зависимостей. Поэтому в случае сложных данных и нелинейных связей эти методы могут оказаться менее эффективными по сравнению с другими алгоритмами, такими как деревья решений или нейронные сети.

# Метод наивного Байеса

# Описание метода

Метод наивного Байеса основан на теореме Байеса и предполагает независимость между признаками. Он использует вероятностный подход к классификации и основывается на простой предпосылке, что признаки объектов независимы между собой при условии принадлежности к определенному классу.

# Применение на наборе данных

На рисунке 8-9 представлены создание модели наивного Байеса и вывод результатов

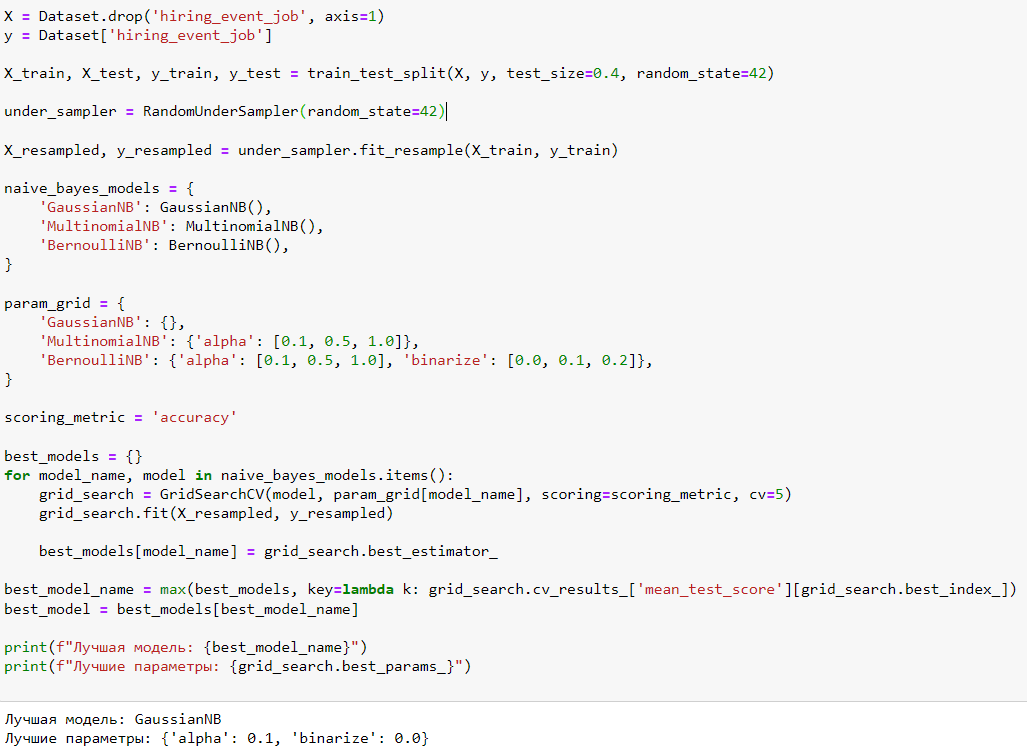


Рис. 8 – Создание модели наивного Байеса и вывод результатов

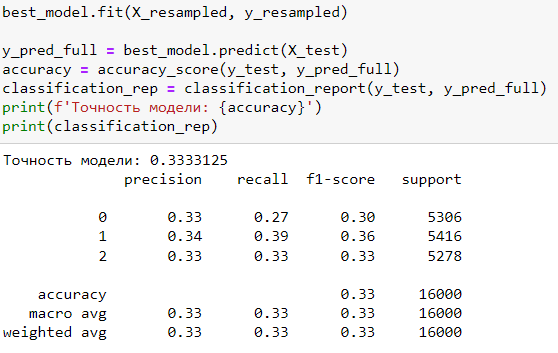


Рис. 9 – Создание модели наивного Байеса и вывод результатов

# Преимущества и ограничения

Преимущества метода наивного Байеса:

* Эффективен и быстр в обучении, особенно при работе с большими наборами данных.
* Хорошо работает в условиях небольшой обучающей выборки.
* Прост в реализации и понимании.

Однако у метода наивного Байеса есть и ограничения:

* Предполагает независимость признаков, что может быть не всегда верным в реальных данных.
* Может давать неоптимальные результаты в случае, когда зависимости между признаками сильны.
  1. **Методы решающего дерева и случайного леса**

# Описание метода

Решающее дерево представляет собой структуру, состоящую из узлов и листьев, которая используется для принятия решений на основе вопросов о значениях признаков. Оно строит дерево, разбивая данные на подмножества на основе определенных признаков.

Принцип работы решающего дерева заключается в выборе наилучшего признака для разделения данных на каждом узле дерева. Этот процесс повторяется, пока не будет достигнуто условие остановки, такое как максимальная глубина дерева или минимальное количество объектов в листе.

# Применение на модельных данных

На рисунке 10 представлен подбор параметров и обучение методом решающего дерева

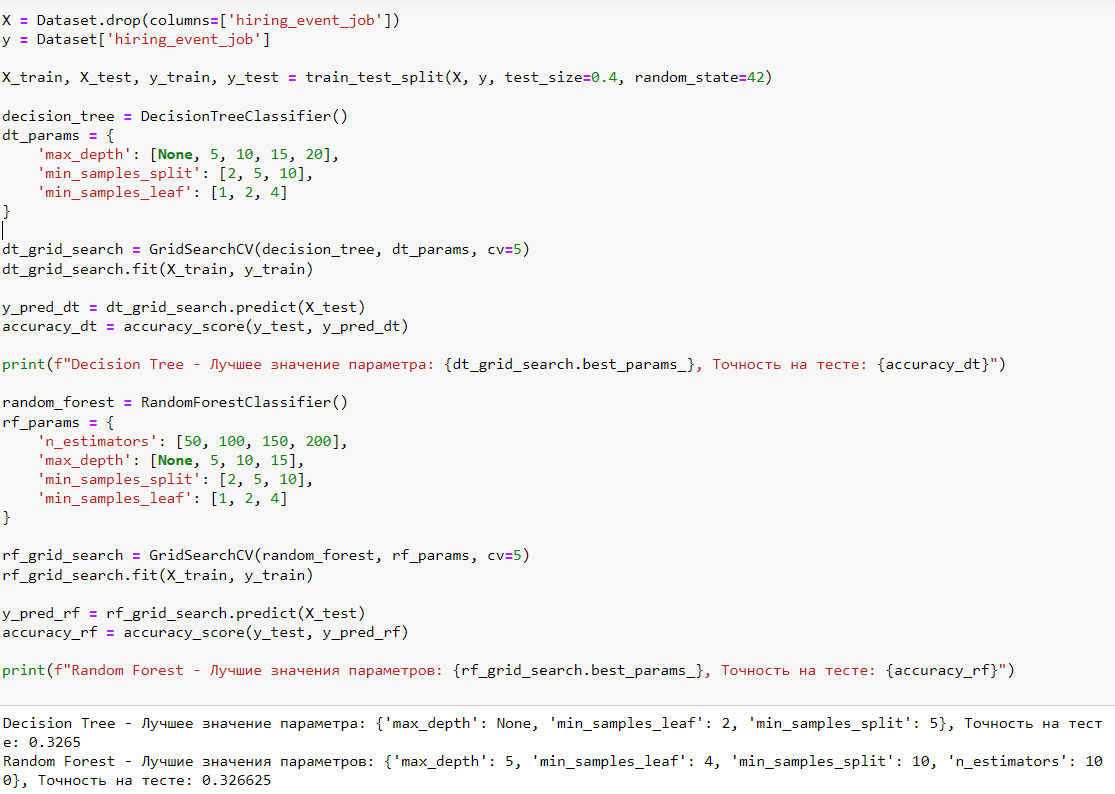


Рис. 10 – Метод решающего дерева

На рисунке 11 представлены обучение решающего дерева и вывод результата

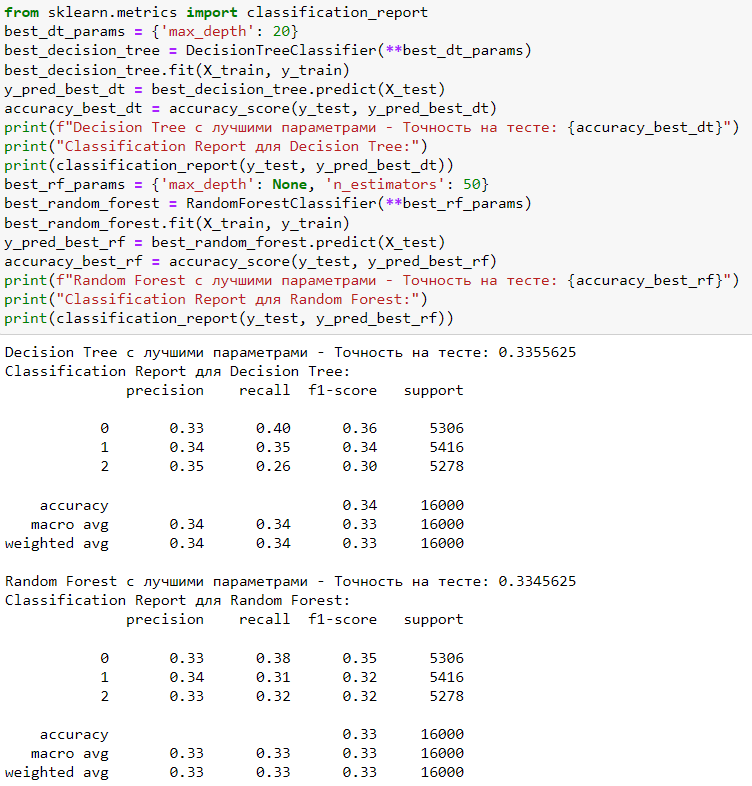


Рис. 11 – Обучение и вывод результата

# Сравнение с другими методами классификации

Решающее дерево и случайный лес являются эффективными методами классификации, которые обладают рядом преимуществ:

* Способны обрабатывать как категориальные, так и числовые данные.
* Позволяют визуализировать принятие решений.
* Могут обрабатывать большие объемы данных.

Однако у этих методов также есть некоторые ограничения:

* Склонны к переобучению на обучающих данных.
* Могут не улавливать сложные нелинейные зависимости.

По сравнению с линейными моделями, решающие деревья и случайный лес могут быть более гибкими в обработке сложных данных, но требуют осторожной настройки параметров для предотвращения переобучения. Их эффективность часто зависит от природы данных и особенностей задачи классификации.

# Метод CatBoost

# Описание метода

CatBoost (Categorical Boosting) – это высокоэффективная библиотека градиентного бустинга, специально разработанная для работы с категориальными признаками. Она представляет собой мощный алгоритм машинного обучения, который широко применяется в задачах классификации, регрессии и ранжирования.

# Применение на модельных данных

На рисунке 12 представлена реализация метода CatBoost

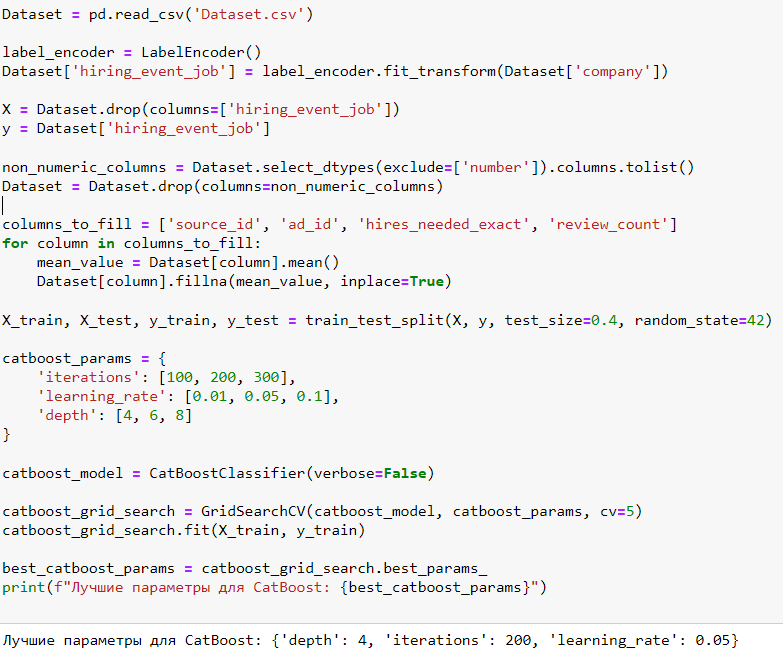


Рис.12 – Метод CatBoost

На рисунке 13 преставлен вывод результата работы метода CatBoost

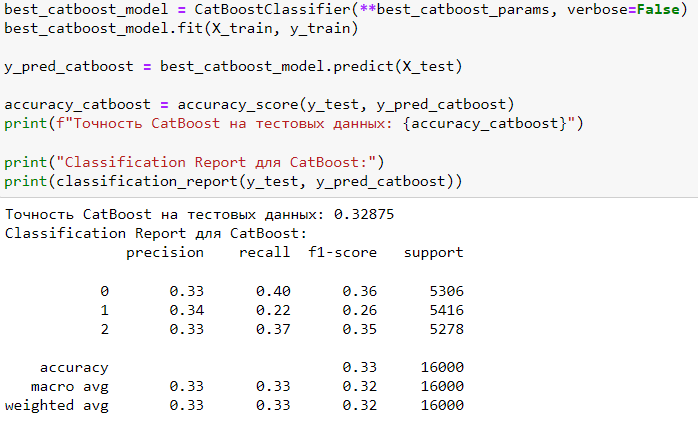


Рис.13 – Вывод результата

CatBoost обладает мощными возможностями для работы с категориальными данными и является одним из популярных алгоритмов для задач машинного обучения, особенно в случаях, когда требуется обработка сложных и разнородных данных без предварительной подготовки.

# Заключение

При выборе метода обучения для модели анализа кредитных заявок или оценки риска мошенничества в банковской сфере, необходимо учитывать не только характеристики датасета, но и цели моделирования. В данном случае, у нас есть набор признаков, описывающих заявителей, и целевой столбец, указывающий, является ли заявка мошеннической или нет.

Из рассмотренных методов машинного обучения следует рассмотреть несколько критериев для выбора подходящего метода:

Размер датасета: Поскольку у вас большой объем данных, методы, такие как случайный лес и градиентный бустинг (например, CatBoost), обычно хорошо справляются с большими объемами данных.

Дисбаланс классов: У вас уменьшили объем датасета и увеличили процент положительных результатов для целевого столбца. Это хороший подход для работы с дисбалансом классов, но также стоит учитывать, что не все методы хорошо справляются с несбалансированными данными.

Интерпретируемость: Если важно понимать, какие признаки оказывают влияние на решение модели, линейные методы (линейная регрессия, логистическая регрессия) могут быть более интерпретируемыми.

Скорость обучения и предсказания: Некоторые методы, такие как k-ближайших соседей, могут быть более вычислительно затратными, особенно с увеличением размера датасета.

Обработка категориальных признаков: Если у вас есть категориальные признаки, CatBoost способен обрабатывать их без предварительного кодирования.

С учетом вышеперечисленных критериев, для модели анализа кредитных заявок рекомендуется рассмотреть методы, такие как CatBoost, случайный лес или градиентный бустинг. Эти методы обладают хорошей производительностью на больших датасетах, способны работать с дисбалансом классов и могут обрабатывать категориальные признаки.

Однако, перед окончательным выбором метода, рекомендуется провести дополнительные эксперименты, включая настройку гиперпараметров и валидацию модели на отложенной выборке, чтобы оценить ее производительность на конкретной задаче и датасете.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Основы машинного обучения, лекция 2 — метод k ближайших соседей: [https://www.youtube.com/watch?v=X081VuXB1og&list=PLEwK9wdS5g0oCR](https://www.youtube.com/watch?v=X081VuXB1og&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=2) [xBzxsq9lkJkzMgzWiyg&index=2](https://www.youtube.com/watch?v=X081VuXB1og&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=2)
2. Основы машинного обучения, лекция 4 — линейная регрессия: [https://www.youtube.com/watch?v=8RAXDT\_5\_js&list=PLEwK9wdS5g0oCR](https://www.youtube.com/watch?v=8RAXDT_5_js&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=24) [xBzxsq9lkJkzMgzWiyg&index=24](https://www.youtube.com/watch?v=8RAXDT_5_js&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=24)
3. Андрей Бурков. Машинное обучение без лишних слов