**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное автономное образовательное  
 учреждение высшего образования   
«Южный федеральный университет»**

**Изображение выглядит как текст, логотип, Шрифт, символ

Автоматически созданное описаниеИнститут высоких технологий и пьезотехники**

**Большие данные**

**Определение мошеннических транзакций**

Выполнили студенты 3 курса 21ВТ-09.03.03.01-о3 группы:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Киливник В.В.

Сим А.Е.

подпись

Проверил старший преподаватель:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Турлюн А. С.

подпись

**Ростов-на-Дону – 2024**

Оглавление

[Введение 3](#_Toc169266366)

[Основная часть 4](#_Toc169266367)

[Графики для анализа зависимостей 11](#_Toc169266368)

[Тренировочные данные. 11](#_Toc169266369)

[Тестовая выборка 12](#_Toc169266370)

[Приступили к обучению модели на методах: Logistic Regression, LinerSVC, Decision tree, Random Forest, Gradient Boosting. 15](#_Toc169266371)

[Логистическая регрессия (Logistic Regression) 15](#_Toc169266372)

[LinearSVC (Linear Support Vector Classification) 16](#_Toc169266373)

[Случайный лес (Random Forest) 18](#_Toc169266374)

[Gradient Boosting (градиентный бустинг) 19](#_Toc169266375)

[ROC кривые 21](#_Toc169266376)

[Тепловая карта 22](#_Toc169266377)

[Кривая точность полнота 23](#_Toc169266378)

[Logistic Roc и Тепловая карта, точность полнота 24](#_Toc169266379)

[DTC Roc и Тепловая карта, точность полнота 28](#_Toc169266380)

[RFC Roc и Тепловая карта, точность полнота 30](#_Toc169266381)

[GBT Roc и Тепловая карта, точность полнота 32](#_Toc169266382)

[Заключение 35](#_Toc169266383)

# Введение

В современном мире онлайн-транзакции стали неотъемлемой частью нашей жизни. С ростом популярности электронных платежей и интернет-магазинов, мошенники все чаще используют разнообразные методы для получения незаконной прибыли. Традиционные методы обнаружения мошенничества часто не справляются с быстро меняющейся тактикой злоумышленников.

Данный проект направлен на разработку системы раннего обнаружения мошеннических транзакций с использованием методов машинного обучения. В основе проекта лежит анализ огромных массивов данных о транзакциях, с целью выявления аномалий и прогнозирования мошеннических действий.

Цель проекта:

* Разработать алгоритм машинного обучения, способный с высокой точностью классифицировать транзакции как мошеннические или легитимные.
* Повысить эффективность обнаружения мошенничества и снизить финансовые потери для банков, платежных систем и других организаций.
* Обеспечить более безопасную и надежную среду для проведения онлайн-транзакций.

Проект будет включать в себя следующие этапы:

* Сбор и подготовка данных: Анализ реальных данных о транзакциях, очистка данных, формирование обучающего и тестового наборов.
* Выбор и обучение моделей машинного обучения.
* Оценка и оптимизация моделей: Проверка точности предсказаний моделей, определение оптимальных параметров, подбор наиболее подходящего алгоритма.

# Основная часть

Для реализации проекта использовали HDFS, Spark и SparkML, MatPlotlib. Seaborn.

Использовали датасет “Credit Card Transactions Fraud Detection Dataset”.

Информация о столбцах датасета:

* index — уникальный идентификатор для каждой строки.
* trans\_date\_trans\_time — Дата и время транзакции
* cc\_num — номер кредитной карты клиента
* торговец - Имя продавца
* категория - Категория продавца
* amt - Сумма транзакции
* first - Имя держателя кредитной карты
* Last – фамилия владельца кредитной карты.
* гендер – пол владельца кредитной карты.
* street – адрес владельца кредитной карты.
* город - Город держателя кредитной карты
* штат - штат держателя кредитной карты
* zip — застежка-молния на держателе кредитной карты
* lat - широта местоположения держателя кредитной карты
* long — долгота местоположения держателя кредитной карты.
* city\_pop — Население города владельца кредитной карты
* вакансия - Вакансия держателя кредитной карты
* dob — дата рождения владельца кредитной карты
* trans\_num - Номер транзакции
* unix\_time - UNIX-время транзакции
* merch\_lat — широта местоположения продавца
* merch\_long — долгота местоположения торговца
* is\_fraud — Флаг мошенничества <--- Целевой класс

Загрузили датасеты в HDFS c помощью команды hdfs dfs -put.



Изображение выглядит как текст, программное обеспечение, число, Значок на компьютере

Автоматически созданное описание

Приступили к написанию кода:

1) В самом начале импортировали нужные библиотеки. Некоторые модули импортировались по ходу написания кода.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

2) Создали функции, которые загружают данные, делают предобработку и обучают модель на различных методах. Данные загружаются из HDFS и затем проводится предобработка: ненужные столбцы удаляются, нулевые значения тоже, категориальные признаки индексируются. Затем создается вектор для обучения модели.

**Описание методов:**

load\_data(path\_to\_archive) – метод, который отвечает за загрузку данных и удаления ненужных столбцов. Возвращает Data Frame.

get\_indexation(df) – метод, который отвечает за индексацию категориальных признаков. Возвращает Data Frame.

mix\_data(df, limit\_1, limit\_0) – метод, который нужен чтою контролировать количество мошеннических транзакций и обычных. (из-за того что в выборке оказалось около 5 процентов мошеннических транзакций пришлось урезать выборку)

generate\_cat\_to\_binary\_vector(df) – преобразование в бинарные векторы категориальных признаков.

Collect\_signs\_to\_vector(df) – собираем все в один вектор.

Evaluator(predictions) – метод для оценки наших моделей.

lern\_logistic\_reg(trainingData): - логистическая модель.

lern\_linear\_SVC(trainingData): - метод опорных векторов.

lern\_DTC(trainingData): - метод классификатор дерева решений.

lern\_RFC(trainingData): - метод случайного леса.

lern\_GBT(trainingData): - метод градиентного дерева.

get\_prediction(classifier, test\_data): - метод для вызова предсказания.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

3) Применили их.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

3) Вывели график количества мошеннических и легальных транзакций в датасете fraudTrain (это именно тренировочные данные).

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

4) Вывели график количества мошеннических и легальных транзакций в датасете fraudTest (данные на которых будет оценка).

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

# Графики для анализа зависимостей

## Тренировочные данные.

1) График зависимости категорий и суммы.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

2) График зависимости суммы от работы

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

3) График зависимости и разбиение на мошеннические и легальные.

Легальные – синий цвет.

Нелегальные – оранжевый.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, программное обеспечение, Графическое программное обеспечение

Автоматически созданное описание

## Тестовая выборка

1) График зависимости категорий и суммы.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

2) График зависимости суммы от работы

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Автоматически созданное описание

3) График зависимости и разбиение на мошеннические и легальные.

Легальные – синий цвет.

Нелегальные – оранжевый.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

# Приступили к обучению модели на методах: Logistic Regression, LinerSVC, Decision tree, Random Forest, Gradient Boosting.

Подробнее о каждом методе:

# Логистическая регрессия (Logistic Regression)

**Принцип:** Логистическая регрессия использует линейную модель для предсказания вероятности принадлежности объекта к определенному классу.

**Как работает:**

* Строит линейную комбинацию предикторов, используя коэффициенты, которые оптимизируются для минимизации ошибки классификации.
* Применяет логистическую функцию (сигмоиду) к линейной комбинации, чтобы получить вероятность принадлежности к классу.

**Преимущества:**

* Простота реализации и интерпретации.
* Работает быстро и эффективно на больших наборах данных.

**Недостатки:**

* Предполагает линейную зависимость между предикторами и целевой переменной.
* Может быть чувствительна к выбросам.

# LinearSVC (Linear Support Vector Classification)

**Принцип:** LinearSVC строит гиперплоскость, которая разделяет данные на классы с максимальным расстоянием от ближайших точек (опорных векторов).

Как работает:

* Ищет оптимальную гиперплоскость, которая минимизирует ошибку классификации и максимизирует расстояние между опорными векторами.

**Преимущества:**

* Высокая точность классификации.
* Робастность к выбросам.

**Недостатки:**

* Может быть сложнее в интерпретации, чем логистическая регрессия.
* Может быть медленнее, чем логистическая регрессия на больших наборах данных.

**Дерево решений (Decision Tree)**

**Принцип:** Дерево решений представляет собой древовидную структуру, которая принимает решения на основе последовательных условий.

**Как работает:**

* Строит дерево, разделяя данные на подгруппы с помощью предикторов, пока не достигается заданное количество подгрупп или пока не удовлетворяются заданные критерии.

**Преимущества:**

* Простота интерпретации.
* Может обрабатывать нелинейные зависимости между предикторами и целевой переменной.

**Недостатки:**

* Может быть склонно к переобучению, особенно на малых наборах данных.
* Может быть чувствительным к шуму в данных.

# Случайный лес (Random Forest)

**Принцип:** Случайный лес состоит из множества деревьев решений, которые обучены на случайных подвыборках данных.

**Как работает:**

* Строит множество деревьев решений, используя различные подвыборки данных и случайные подмножества предикторов.
* Использует голосование (или усреднение) по всем деревьям, чтобы получить окончательный прогноз.

**Преимущества:**

* Высокая точность классификации.
* Робастность к выбросам и переобучению.

**Недостатки:**

* Может быть сложнее в интерпретации, чем дерево решений.
* Может быть медленнее, чем дерево решений на больших наборах данных.

# Gradient Boosting (градиентный бустинг)

**Gradient Boosting (градиентный бустинг)** — это алгоритм машинного обучения, который объединяет множество слабых моделей, называемых деревьями решений, в сильную модель. Он работает по принципу последовательного обучения:

1. Инициализация: Строится начальная модель, например, среднее значение целевой переменной.

2. Обучение:

- Для каждого обучающего примера рассчитывается ошибка предсказания текущей модели.

- Обучается слабое дерево решений, минимизирующее ошибку предсказания.

- Предсказания дерева добавляются к текущему предсказанию модели, с весом, пропорциональным точности дерева.

3. Повторение: Шаги 2-3 повторяются до достижения желаемой точности модели.

**Ключевые особенности Gradient Boosting:**

Последовательное обучение: Каждое дерево обучается на основе ошибок предыдущих деревьев.

Градиентный спуск: Алгоритм использует градиентный спуск для оптимизации параметров модели.

Слабые деревья решений: Используются простые модели деревьев решений, которые могут быть легко интерпретированы.

Регуляризация: Используются техники регуляризации для предотвращения переобучения.

**Преимущества Gradient Boosting:**

Высокая точность: Gradient Boosting обычно демонстрирует высокую точность предсказания.

Устойчивость к шуму: Алгоритм устойчив к шуму в данных.

Интерпретируемость: Хотя он использует множество деревьев решений, его предсказания могут быть интерпретированы с помощью анализа важности признаков.

**Недостатки Gradient Boosting:**

Сложность настройки: Требуется тщательная настройка параметров алгоритма для достижения оптимальной производительности.

Высокая вычислительная стоимость: Gradient Boosting может быть вычислительно дорогим для больших наборов данных.

Чрезмерное переобучение: Алгоритм может переобучиться, если не использовать правильную регуляризацию.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

7) Вывели графики ROC кривых и прочее для оценки качества обучения модели.

ROC-кривая — график, позволяющий оценить качество [бинарной классификации](https://ru.wikipedia.org/wiki/Бинарная_классификация), отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак при варьировании порога решающего правила.

Самый лучший результат показал метод Gradient Boosting.

# ROC кривые

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

# Тепловая карта

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

# Кривая точность полнота

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

## Logistic Roc и Тепловая карта, точность полнота

Точность модели: 84

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

SVC Roc и Тепловая карта, точность полнота

Точность модели: 84

Изображение выглядит как линия, текст, снимок экрана, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

## DTC Roc и Тепловая карта, точность полнота

Точность модели: 89

Изображение выглядит как линия, текст, снимок экрана, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, дизайн

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, дисплей, линия

Автоматически созданное описание

## RFC Roc и Тепловая карта, точность полнота

Точность модели: 91

Изображение выглядит как линия, снимок экрана, текст, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

## GBT Roc и Тепловая карта, точность полнота

Точность модели: 97

Изображение выглядит как линия, текст, снимок экрана, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, дизайн

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, дисплей, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

# Заключение

В ходе проекта были достигнуты следующие ключевые результаты:

* Разработан алгоритм машинного обучения, способный с высокой точностью классифицировать транзакции как мошеннические или легитимные.
* Проведена оптимизация алгоритма, обеспечивающая максимальную точность предсказаний и минимизацию ложных срабатываний.
* Создана система раннего оповещения, интегрирующая алгоритм машинного обучения в систему мониторинга транзакций.
* Проведено тестирование системы, подтвердившее ее эффективность в обнаружении мошеннических действий.

В результате проекта была разработана система, способная значительно улучшить защиту от мошенничества в сфере онлайн-платежей. Полученные результаты демонстрируют потенциал машинного обучения в повышении безопасности онлайн-транзакций и снижении финансовых потерь для пользователей и организаций.

Проект является отличным примером того, как машинное обучение может быть использовано для решения реальных проблем в современном мире. Он открывает новые перспективы в области безопасности онлайн-платежей и позволяет создать более безопасную и надежную среду для всех участников финансовых транзакций.