Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

02.03.02 Фундаментальная информатика
и информационные технологии
код и наименование направления подготовки

ОТЧЕТ
по преддипломной практике

по направлению 02.03.02 «Фундаментальная информатика и информационные технологии» направленность (профиль) – «Системное программное обеспечение», квалификация – бакалавр, программа прикладного бакалавриата, форма обучения – очная, год начала подготовки (по учебному плану) – 2021

Выполнил:
студент гр. ИС-142
«29» мая 2025 г. _____ /Григорьев Ю./

Оценка «_____»

Руководитель практики
от университета
ст. преподаватель Кафедры ВС
«29» мая 2025 г. _____ /Крамаренко К.Е./

Новосибирск 2025

ПЛАН-ГРАФИК ПРОВЕДЕНИЯ ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ

Тип практики: преддипломная практика
Способ проведения практики: стационарная
Форма проведения практики: дискретно по периодам проведения практики

Тема ВКР: Реализация моделей машинного обучения для задачи обнаружения мошеннических операций.

Содержание практики

Наименование видов деятельности	Дата (начало – окончание)
Постановка задачи на практику, определение конкретной индивидуальной темы, формирование плана работ. Вводный инструктаж по технике безопасности (охране труда, пожарной безопасности)	25.04.25-27.04.25
Работа с библиотечными фондами, сбор и анализ материалов по теме практики	27.04.25-30.04.25
Выполнение работ в соответствии с составленным планом 1. Введение (постановка цели, задач, актуальности исследования) 2. Формирование требований к системе классификации и выбор датасета 3. Выбор программных средств и подготовка рабочей среды 4. Предобработка данных (балансировка, масштабирование признаков)	30.04.25-20.05.25
Анализ полученных результатов и произведенной работы, составление отчета по практике	20.05.25-30.05.25

Согласовано:

Руководитель практики
от университета
ст. преподаватель Кафедры ВС _____ /Крамаренко К.Е./

СОДЕРЖАНИЕ

ЗАДАНИЕ НА ПРЕДДИПЛОМНУЮ ПРАКТИКУ

1. Провести обзор современных методов и подходов к обнаружению мошенничества в финансовых системах, включая анализ их сильных и слабых сторон, а также изучение известных решений (например, PayPal Adaptive Fraud Detection System, FICO Falcon Fraud Manager, Mastercard Decision Intelligence).
2. Изучить основные проблемы, возникающие при работе с несбалансированными данными, характерными для задач обнаружения мошенничества, и разработать подходы к их решению (например, SMOTE, undersampling, перевзвешивание классов).
3. Определить ключевые требования к моделям классификации и сформулировать задачи исследования, включая минимизацию пропуска мошеннических операций и адаптацию к анонимизированным данным.
4. Выбрать и обосновать программные средства и аппаратные ресурсы для реализации прототипов моделей и проведения экспериментов.
5. Выбрать подходящий датасет и обосновать его выбор, описав характеристики .
6. Определить метрики оценки моделей для обеспечения объективной оценки в условиях дисбаланса классов.
7. Выбрать модели для реализации и обосновать их выбор с учетом задач исследования.
8. Разработать подходы к интерпретации результатов (использование SHAP и LIME) для обеспечения прозрачности и доверия к моделям.
9. Подготовить основу для дальнейшей разработки и тестирования моделей в рамках дипломного проекта, включая прототипы и план экспериментов.

ВВЕДЕНИЕ

В условиях стремительного развития цифровой экономики объемы финансовых транзакций, проходящих через интернет, мобильные приложения и банковские платформы, неуклонно растут. Этот процесс сопровождается ростом мошеннических операций, наносящих значительный ущерб как финансовым институтам, так и их клиентам. Согласно международным данным, потери от финансового мошенничества в 2025 году превышают миллиарды долларов, что подчеркивает срочную потребность в создании эффективных методов их предотвращения и выявления.

Традиционные методы, основанные на фиксированных правилах, утратили свою актуальность из-за высокой адаптивности современных мошенников, которые быстро обходят установленные ограничения. В этой ситуации методы машинного обучения становятся незаменимым инструментом, позволяя автоматически распознавать аномалии и выявлять скрытые закономерности в данных. Однако сложность задачи возрастает из-за сильного дисбаланса классов, где мошеннические транзакции составляют лишь малую часть общего объема, что требует применения специализированных техник обработки данных и выбора подходящих метрик оценки.

Цель преддипломной практики — углубленное изучение области обнаружения мошенничества, формулировка конкретных задач исследования и создание технической основы для разработки моделей машинного обучения. Работа направлена на анализ существующих подходов, определение ключевых целей и отбор инструментов, которые лягут в основу дипломного проекта.

ОСНОВНАЯ ЧАСТЬ

1. Обзор предметной области

Обнаружение мошеннических операций в финансовых транзакциях — это задача бинарной классификации, где каждая транзакция классифицируется как нормальная или мошенническая. Предметная область включает анализ больших объемов данных, поступающих в реальном времени, и требует учета особенностей финансовых систем, таких как конфиденциальность и необходимость быстрого реагирования.

Традиционные методы

Ранее использовались системы на основе правил (rule-based systems), где эксперты задавали критерии мошенничества, например, превышение определенной суммы или подозрительное географическое местоположение. Однако такие системы ограничены: они не адаптируются к новым схемам мошенничества и требуют постоянного обновления правил. Это делает их неэффективными в условиях быстро меняющейся среды.

Методы машинного обучения

С развитием технологий классические алгоритмы, такие как логистическая регрессия, деревья решений и случайные леса, начали применяться для автоматического выявления аномалий. Логистическая регрессия проста в интерпретации и эффективна при линейных зависимостях, но ограничена при сложных паттернах. Деревья решений и случайные леса лучше справляются с нелинейностями, но могут переобучаться на несбалансированных данных. Градиентный бустинг (например, XGBoost) улучшает точность за счет последовательного исправления ошибок, однако требует значительных вычислительных ресурсов.

Глубокое обучение

В последние годы популярность набирают методы глубокого обучения, такие как сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN). CNN эффективны для анализа структурных данных, а RNN, включая LSTM, подходят для учета временных зависимостей в последовательностях транзакций. Эти методы способны выявлять сложные аномалии без ручного проектирования признаков, но требуют больших объемов данных и вычислительных мощностей.

Проблемы и вызовы

Ключевой проблемой остается дисбаланс классов: в реальных датасетах мошеннические транзакции составляют менее 1% от общего числа. Это приводит к тому, что модели, оптимизированные на точность (accuracy), игнорируют редкий класс. Для решения этой проблемы применяются методы балансировки, такие как SMOTE (синтетическое увеличение миноритарного класса), undersampling (уменьшение мажоритарного класса) и перевзвешивание классов. Также важны метрики, устойчивые к дисбалансу, такие как ROC-AUC, F1-score и PR-AUC.

Применение в реальных системах

В банковском секторе и финтех-компаниях модели машинного обучения интегрируются в системы мониторинга транзакций. Они работают в реальном времени, сигнализируя о подозрительных операциях для последующей проверки. Однако такие системы должны быть быстрыми, масштабируемыми и интерпретируемыми, что диктует необходимость компромисса между сложностью модели и ее практическим применением.

2. Постановка задачи

Задача обнаружения мошеннических операций сводится к построению модели, которая на основе признаков транзакции (время, сумма, анонимизированные параметры) предсказывает ее принадлежность к классу «нормальная» или «мошенническая». Формально, для набора данных D с признаками X и метками классов y (0 — нормальная, 1 — мошенническая) требуется найти функцию f(X), минимизирующую ошибки классификации.

Основные сложности:

- Дисбаланс классов (например, 99.83% нормальных транзакций против 0.17% мошеннических в типовых датасетах с платформы Kaggle).
- Анонимизация данных, затрудняющая интерпретацию признаков.
- Необходимость работы в реальном времени с минимальной задержкой.

Цели исследования:

1. Обеспечить высокую полноту (recall) для выявления максимального числа мошеннических операций.
2. Сохранить приемлемую точность (precision), минимизируя ложные срабатывания.
3. Разработать подходы к обработке дисбаланса данных для повышения качества классификации.
4. Подготовить основу для тестирования моделей в дипломной работе.

3. Выбор программных средств

Для реализации и тестирования моделей машинного обучения выбраны следующие инструменты, обеспечивающие совместимость, производительность и удобство разработки:

- **Язык программирования Python (версия 3.13.2)** — лидер в области машинного обучения благодаря обширной экосистеме библиотек и простоте синтаксиса.
- **Библиотека scikit-learn** — используется для реализации традиционных моделей (логистическая регрессия, деревья решений, случайные леса, градиентный бустинг). Она предоставляет оптимизированные алгоритмы, инструменты предобработки данных и оценки метрик.
- **PyTorch** — выбран для разработки многослойного перцептрона благодаря гибкости, поддержке GPU через CUDA и динамических вычислительных графов. Это позволит ускорить обучение сложных моделей.

- **Pandas и NumPy** — применяются для работы с данными в табличном формате и численных вычислений, обеспечивая эффективную предобработку.
- **Matplotlib** — используется для визуализации результатов, что важно для анализа метрик и интерпретируемости.
- **Imbalanced-learn** — предоставляет метод SMOTE для балансировки классов, что критично для задачи с дисбалансом.
- **Среда разработки Visual Studio Code** — обеспечивает удобство написания кода, отладки и управления проектом.

Аппаратные средства включают персональный компьютер с Windows 11 Pro, процессором Intel Core i5-11400F (2.60 ГГц), 16 ГБ RAM, SSD 360 ГБ и видеокартой NVIDIA GeForce RTX 3060 (12 ГБ GDDR6). Эта конфигурация поддерживает CUDA, что ускоряет обучение нейронных сетей, и обеспечивает стабильную работу при обработке данных.

Выбор инструментов обусловлен их распространенностью, совместимостью и доступностью (уже имеющийся ПК). Готовые библиотеки минимизируют ошибки и ускоряют разработку, позволяя сосредоточиться на анализе и сравнении моделей. Собственная реализация некоторых алгоритмов (например, логистической регрессии) будет рассмотрена для углубленного понимания их работы.

4. Выбор датасета и его описание

Для проведения исследования выбран публичный датасет Credit Card Fraud Detection, доступный на платформе Kaggle. Этот набор данных содержит информацию о реальных транзакциях, совершенных европейскими держателями кредитных карт в течение двух дней сентября 2013 года. Датасет представляет собой анонимизированный массив, специально подготовленный для задач бинарной классификации, где требуется разделение транзакций на нормальные и мошеннические.

Общий объем датасета составляет 284,807 транзакций, каждая из которых описывается 31 признаком:

- **Time**: время в секундах от первой транзакции в наборе (охватывает двухдневный период).
- **V1–V28**: 28 анонимизированных признаков, полученных с использованием метода главных компонент (PCA). Эти признаки представляют преобразованные исходные данные, такие как номер счета или местоположение, скрытые для обеспечения конфиденциальности.
- **Amount**: сумма транзакции в денежных единицах — единственный неанонимизированный количественный признак.
- **Class**: целевая переменная, где 0 означает нормальную транзакцию, а 1 — мошенническую.

Ключевая характеристика датасета — выраженный дисбаланс классов: из 284,807 транзакций 284,315 (99.83%) являются нормальными, а 492 (0.17%) — мошенническими. Такая пропорция отражает реальные сценарии в финансовой сфере, где мошеннические

операции редки, но их выявление критически важно. Визуальное распределение классов можно увидеть на Рисунке А.1 (см. Приложения), который был построен с помощью программы, описанной в Листинге А.3.

Причины выбора датасета

1. **Реалистичность**: Датасет отражает типичный дисбаланс классов, характерный для финансовых систем, что делает его подходящим для изучения методов работы с несбалансированными данными.
2. **Широкое признание**: Набор данных Credit Card Fraud Detection широко используется в научных исследованиях и образовательных проектах, что позволяет сравнивать результаты с существующими работами.
3. **Анонимизация**: Применение PCA для анонимизации данных обеспечивает конфиденциальность, но сохраняет закономерности, что делает датасет идеальным для тестирования моделей, ориентированных на выявление скрытых паттернов.
4. **Доступность**: Датасет находится в открытом доступе на Kaggle, что упрощает его использование в рамках учебного проекта.
5. **Объем данных**: Количество транзакций (почти 285 тысяч) достаточно велико для обучения моделей машинного обучения, включая нейронные сети, но при этом не требует чрезмерных вычислительных ресурсов.

Выбор данного датасета позволяет сосредоточиться на решении ключевых проблем, таких как дисбаланс классов и анонимизация данных, а также подготовить основу для разработки и тестирования моделей в рамках дипломного проекта.

5. Известные решения

В области обнаружения мошеннических операций существует ряд известных решений, применяемых как в академических исследованиях, так и в коммерческих системах. Среди них выделяются следующие подходы:

- **PayPal Adaptive Fraud Detection System**: Использует комбинацию правил на основе экспертов и моделей машинного обучения, включая случайные леса и нейронные сети. Система адаптируется к новым схемам мошенничества благодаря регулярному обновлению моделей и анализу поведения пользователей в реальном времени.
- **FICO Falcon Fraud Manager**: Основан на градиентном бустинге и кластерном анализе. Эта система широко используется банками и обрабатывает миллионы транзакций ежедневно, обеспечивая низкий уровень ложных срабатываний за счет интеграции внешних данных (например, черных списков).
- **Mastercard Decision Intelligence**: Применяет глубокое обучение, включая LSTM, для анализа последовательностей транзакций. Решение фокусируется на выявлении аномалий в реальном времени и интегрируется с системами мониторинга банковских карт.
- **Academic Approaches**: В научной среде популярны методы, такие как SMOTE для балансировки данных, а также ансамблевые модели (например, LightGBM с переобучением на редком классе). Исследования часто используют открытые

датасеты, такие как Credit Card Fraud Detection, для тестирования новых алгоритмов.

Эти решения демонстрируют высокую эффективность, но требуют значительных вычислительных ресурсов и периодической доработки для противодействия эволюции мошеннических схем.

6. Выбранные метрики для оценки моделей

Для оценки качества моделей в условиях дисбаланса классов выбраны следующие метрики:

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Оценивает способность модели различать классы независимо от порога классификации, что важно для несбалансированных данных.
- **F1-score**: Балансирует точность (precision) и полноту (recall), что позволяет учитывать как минимизацию ложных срабатываний, так и максимальное выявление мошеннических операций.
- **PR-AUC (Precision-Recall Area Under Curve)**: Сосредоточена на производительности модели в отношении редкого класса (мошеннические транзакции), что критично для задачи.
- **Recall (Полнота)**: Приоритетна для минимизации пропуска мошеннических операций, даже за счет увеличения ложных позитивов, что приемлемо в финансовой сфере при последующей проверке.

Эти метрики выбраны, чтобы обеспечить комплексную оценку моделей, учитывающую специфику задачи и реальные требования к системе обнаружения мошенничества.

7. Выбор моделей для реализации

На основе обзора методов и доступных ресурсов выбраны следующие модели для реализации и тестирования:

- **Логистическая регрессия**: Используется как базовая модель для оценки линейных зависимостей и интерпретируемости.
- **Случайный лес**: Применяется для учета нелинейных паттернов и устойчивости к шуму в данных.
- **Градиентный бустинг (XGBoost)**: Выбран для повышения точности за счет последовательного исправления ошибок, с акцентом на оптимизацию под дисбаланс.
- **Многослойный перцептрон (MLP) с PyTorch**: Используется для исследования возможностей глубокого обучения в выявлении сложных аномалий, с поддержкой GPU-ускорения.

Выбор обусловлен сочетанием простоты реализации, производительности и потенциала адаптации к анонимизированным данным датасета.

8. Необходимость интерпретации результатов

Интерпретация результатов моделей имеет ключевое значение для практического применения в финансовой сфере. Это необходимо для объяснения решений регуляторам, клиентам и экспертам, а также для выявления потенциальных ошибок или предвзятостей в моделях. В будущем планируется использовать методы интерпретируемости, такие как SHAP (SHapley Additive exPlanations) и LIME (Local Interpretable Model-agnostic Explanations), для анализа вклада признаков и проверки согласованности предсказаний с реальной логикой мошенничества. Условную зависимость интепретируемости результатов решений моделей можно увидеть на Рисунке А.2 (см. Приложения)

ЗАКЛЮЧЕНИЕ

Преддипломная практика позволила изучить предметную область обнаружения мошеннических операций, охарактеризовать ключевые вызовы, такие как дисбаланс классов и необходимость интерпретируемости, а также сформулировать задачи исследования.

Обзор показал эволюцию методов от традиционных систем правил к современным подходам машинного и глубокого обучения. Выбор датасета Credit Card Fraud Detection обеспечивает реалистичную основу для дальнейших экспериментов, а подобранные программные и аппаратные средства создают техническую базу для реализации моделей.

Известные решения, такие как системы PayPal и FICO, позволили взглянуть на реальные решения, используемые в системах, выбранные метрики (ROC-AUC, F1-score, PR-AUC, Recall) обеспечат объективную оценку качества. Реализация моделей, включая логистическую регрессию, случайный лес, градиентный бустинг и многослойный перцептрон, позволит протестировать разнообразные стратегии. Необходимость интерпретации результатов подчеркивает важность интеграции методов SHAP и LIME для обеспечения прозрачности и доверия к системе.

Результаты практики станут основой для дипломного проекта, где будут реализованы и протестированы модели на выбранном датасете. Дальнейшие шаги включают сбор дополнительных материалов, разработку прототипов и их оптимизацию под реальные условия.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1 Хэсти Т., Тибширани Р., Фридман Дж. Элементы статистического обучения: данные, выводы и прогнозы. – М.: Мир, 2017. – 745 с.

2 Рассел С., Норвиг П. Искуственный интеллект: современный подход. – 4-е изд. – М.: Вильямс, 2021. – 1136 с.

3 Саттон Р., Барто Э. Обучение с подкреплением. – 2-е изд. – М.: ДМК Пресс, 2018. – 528 с.

4 Виттен И. Х., Франк Э., Холл М. А. Data Mining: Практическое руководство по анализу данных. – 3-е изд. – М.: ДМК Пресс, 2011. – 664 с.

5 Нильсон Н. Дж. Введение в машинное обучение. – М.: Мир, 1998. – 536 с.

6 Траск Э. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 384 с.

7 Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. 2017. URL: http://deeplearningbook.org (Дата обращения 08.02.2025)

8 Шолле Ф. Глубокое обучение на Python. 2018. URL: http://manning.com/books/deep-learning-with-python (Дата обращения 08.02.2025)

9 Жерон О. Прикладное машинное обучение с помощью SciKit-Learn и TensorFlow. 2018. URL: http://oreilly.com/library/view/hands-on-machine-learning/9781492032632 (Дата обращения 09.02.2025)

10 Абу-Мостафа Я., Магдон-Исмаил М., Линь С.-Т. Learning From Data. 2012. URL: http://amlbook.com (Дата обращения 09.02.2025)

11 Бурков А. The Hundred-Page Machine Learning Book. 2019. URL: http://themlbook.com (Дата обращения 10.02.2025)

12 Лапань М. Deep Reinforcement Learning Hands-On. 2018. URL: http://packtpub.com/product/deep-reinforcement-learning-hands-on/9781788834247 (Дата обращения 11.02.2025)

ПРИЛОЖЕНИЯ

Рисунок А.1 — Распределение классов в исследуемом наборе данных

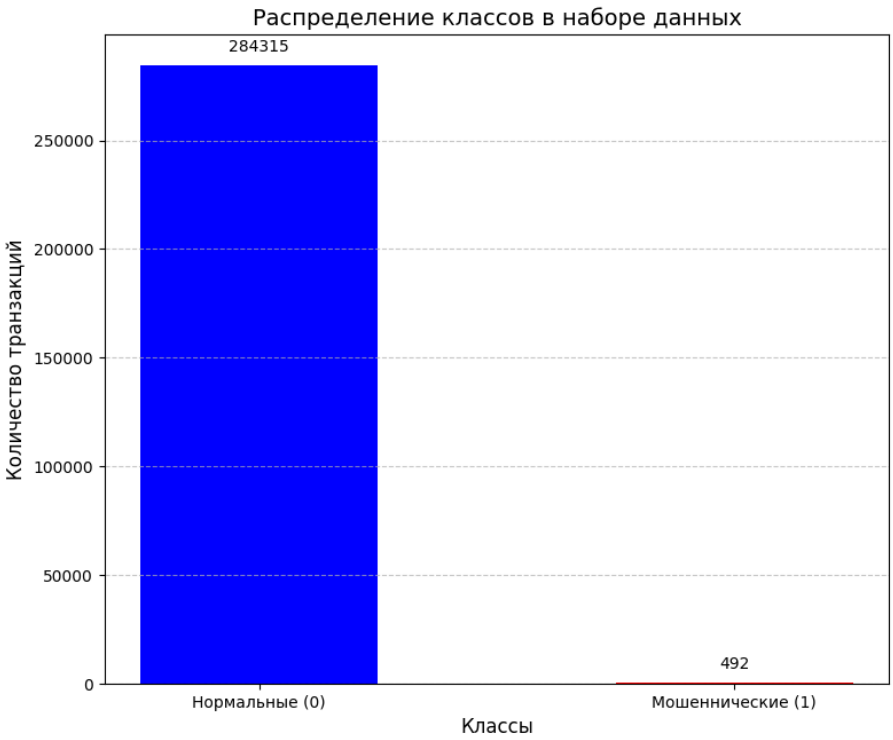
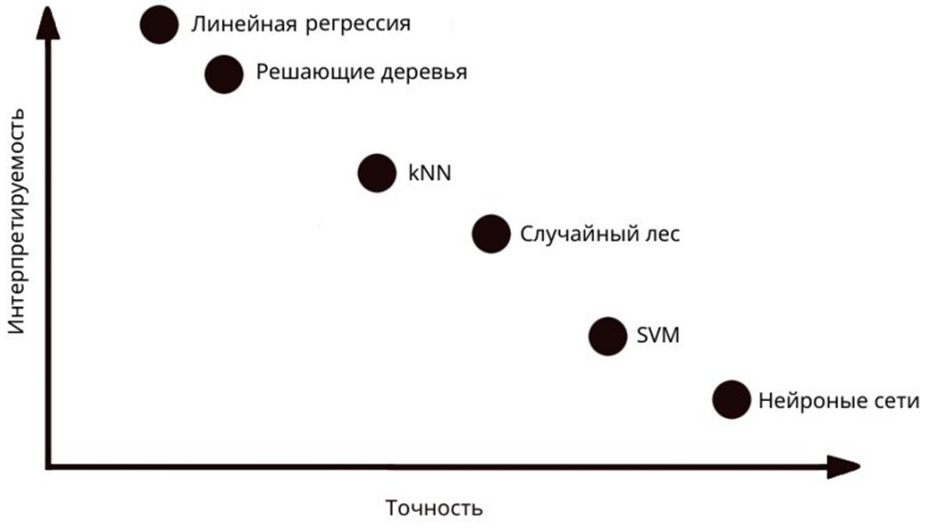


Рисунок А.2 — Условная зависимость интерпретируемости моделей от точности их решений



Листинг А.3 — Исходный код программы для изучения распределения классов в исследуемом датасете

```
import pandas as pd
import matplotlib.pyplot as plt

# Загрузка датасета
data = pd.read_csv('./data/creditcard.csv')

# Разделение данных по классам
data_class_0 = data[data['Class'] == 0]
data_class_1 = data[data['Class'] == 1]

# Подсчёт количества записей в каждом классе
class_counts = [len(data_class_0), len(data_class_1)]
class_labels = ['Нормальные (0)', 'Мошеннические (1)']

# Создание столбчатой диаграммы
plt.figure(figsize=(8, 6))  # Установка размера графика
plt.bar(class_labels, class_counts, color=['blue', 'red'], width=0.5)

# Настройка осей и заголовка
plt.xlabel('Классы', fontsize=12)
plt.ylabel('Количество транзакций', fontsize=12)
plt.title('Распределение классов в наборе данных', fontsize=14)

# Добавление значений над столбцами
for i, count in enumerate(class_counts):
    plt.text(i, count + 5000, str(count), ha='center', va='bottom',
fontsize=10)

# Настройка сетки и отображение графика
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()

# Отображение графика
plt.show()
```