

Выпускная квалификационная работа
**“Прогнозирование прибыли от инвестиционной
деятельности на финансовом рынке”**

по программе профессиональной переподготовки:
«Анализ данных на языке Python»

Выполнил:
Андреев Антон Германович



Актуальность темы

Обзор ситуации:

- Развитие интернет-технологий
- Развитие финтех отрасли
- Доступность инвестиций

Проблема:

необходимость прогнозирования прибыли от инвестиций для принятия правильного решения

Пути решения:

- Экспертная оценка инвестиционного портфеля
- Предоставление брокером информационных материалов
- **Создание data-driven инструмента прогнозирования**

Цель и задачи

Цель работы –

разработка модели прогнозирования прибыли от инвестиций в зависимости от ряда вводных данных, таких как:

- Инвестиционный инструмент
- Степень риска инвестиционного портфеля
- Размер инвестируемого капитала
- Срок инвестиций
- Объем оборота капитала (лотов)

Задачи:

- Сбор внутренних данных об инвестиционной активности клиентов
- Анализ и обработка данных для использования в модели
- Выбор модели прогнозирования
- Реализация инструмента прогнозирования и его тестирования

Требования к инструменту прогнозирования

Безопасность данных

Высокая скорость получения результата

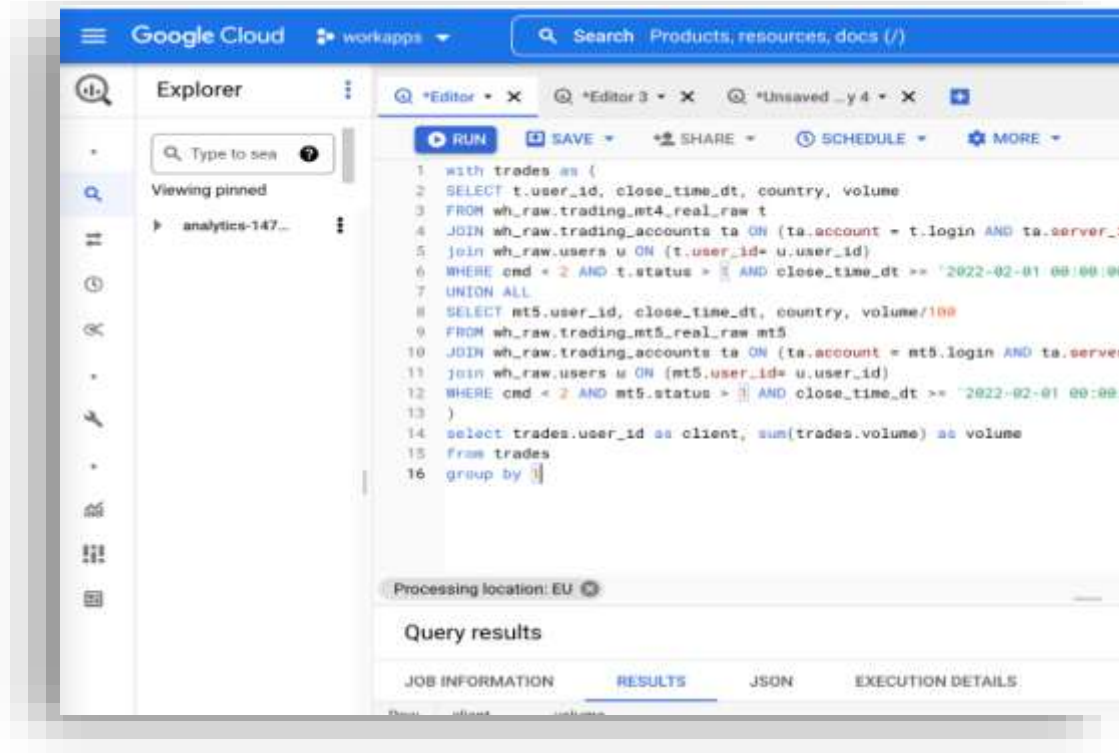
Стабильность работы

Достоверность результатов

Простота реализации

Сбор данных

- Определение перечня необходимых данных
- SQL-запрос
- Загрузка данных на локальную машину
- Объединение нескольких таблиц в один датасет



The screenshot shows the Google Cloud BigQuery interface. The top navigation bar includes the Google Cloud logo, a search bar with the text "Search Products, resources, docs (/)", and a "workapps" dropdown. The main interface is divided into three sections: Explorer, Editor, and Query results.

Explorer: Shows a search bar with "Type to sea" and a "Viewing pinned" section with a folder named "analytics-147...".

Editor: Contains a SQL query with the following code:

```
1 with trades as (  
2 SELECT t.user_id, close_time_dt, country, volume  
3 FROM wh_raw.trading_mt4_real_raw t  
4 JOIN wh_raw.trading_accounts ta ON (ta.account = t.login AND ta.server_1  
5 JOIN wh_raw.users u ON (t.user_id= u.user_id)  
6 WHERE cmd = 2 AND t.status > 0 AND close_time_dt >= '2022-02-01 00:00:00'  
7 UNION ALL  
8 SELECT mt5.user_id, close_time_dt, country, volume/100  
9 FROM wh_raw.trading_mt5_real_raw mt5  
10 JOIN wh_raw.trading_accounts ta ON (ta.account = mt5.login AND ta.server  
11 JOIN wh_raw.users u ON (mt5.user_id= u.user_id)  
12 WHERE cmd = 2 AND mt5.status > 0 AND close_time_dt >= '2022-02-01 00:00:  
13 )  
14 select trades.user_id as client, sum(trades.volume) as volume  
15 from trades  
16 group by client
```

Below the editor, there are buttons for "RUN", "SAVE", "SHARE", "SCHEDULE", and "MORE". The "Processing location" is set to "EU".

Query results: Shows a tabbed interface with "JOB INFORMATION", "RESULTS", "JSON", and "EXECUTION DETAILS". The "RESULTS" tab is active, showing a table with columns "client" and "volume".

Обработка данных

Очистка данных



Модуль `FuzzuWuzzy`, функции

Группировка данных



Метод `.groupby()`, `lambda` функции

Кодирование текстовых значений



Модуль `OrdinalEncoder`, методы `Pandas`

Шифрование данных



Функция шифрования, методы `Pandas`

Нахождение и удаление экстремумов



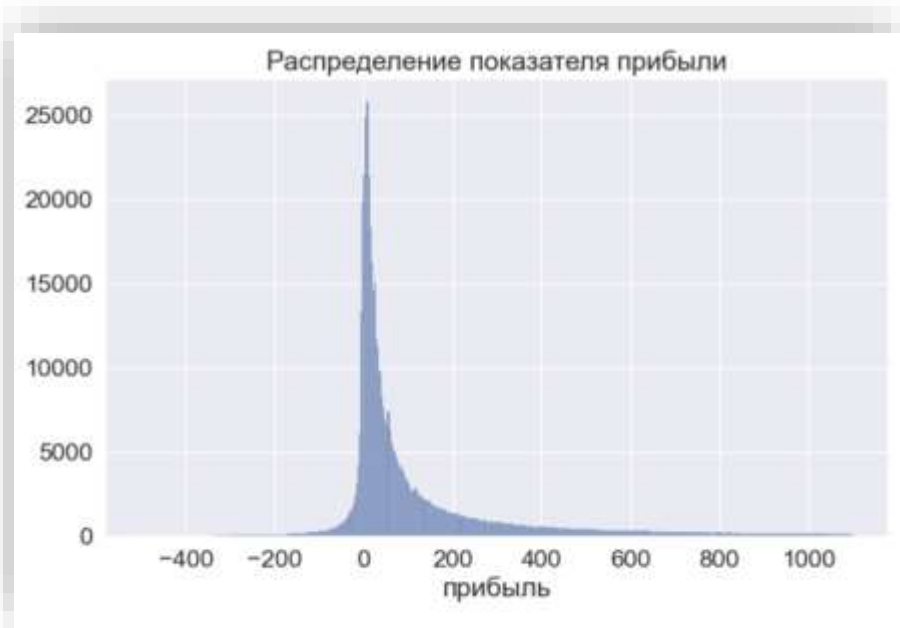
Функции, 90% перцентиль

Первичный анализ данных

```
df.dtypes
```

тип актива	int64
уровень риска портфеля	int64
сумма инвестиций	float64
сумма первого депозита	float64
сумма повторного депозита	float64
инвестиционный период	int64
объем оборота капитала	float64
прибыль	float64
dtype:	object

Анализ типов данных



Гистограмма распределения
результатирующего показателя

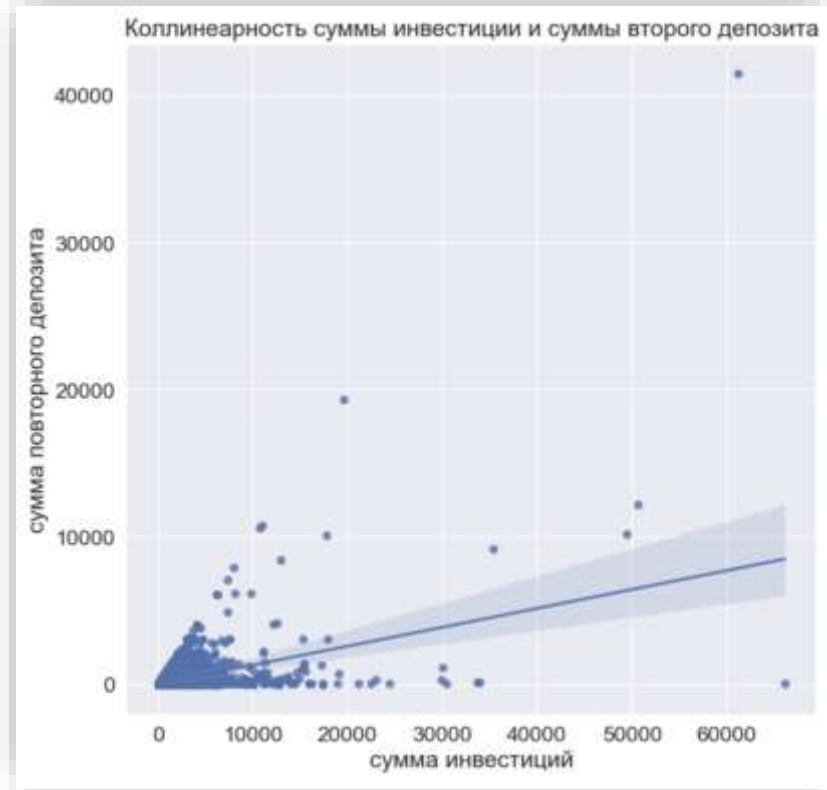
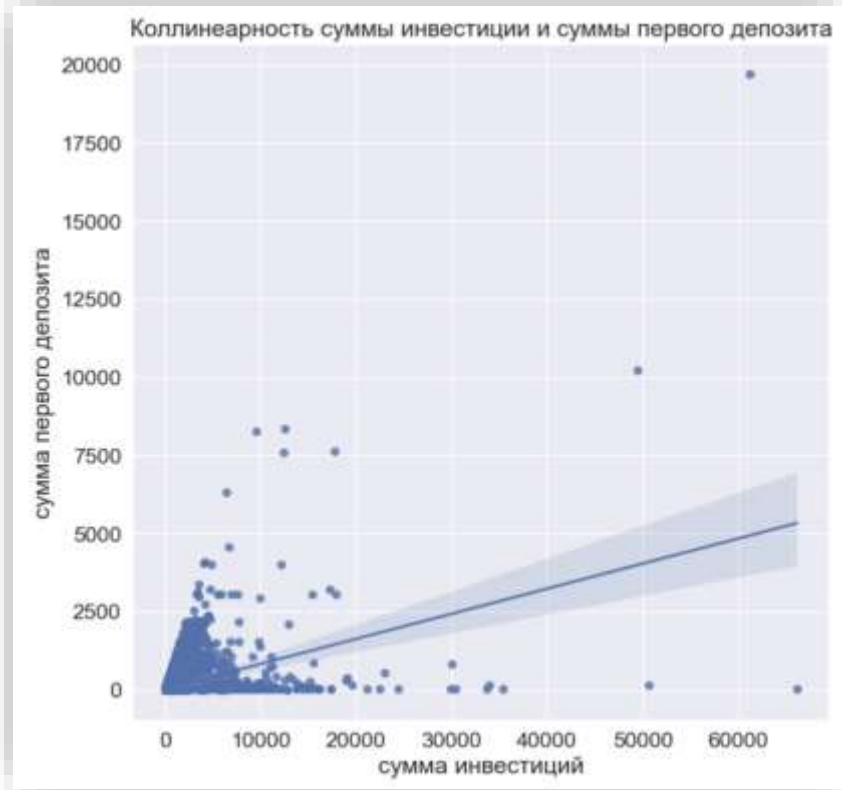
Корреляционный анализ данных

Тепловая карта корреляции показателей



– Тепловая карта корреляции показала возможное наличие коллинеарности не результирующих показателей

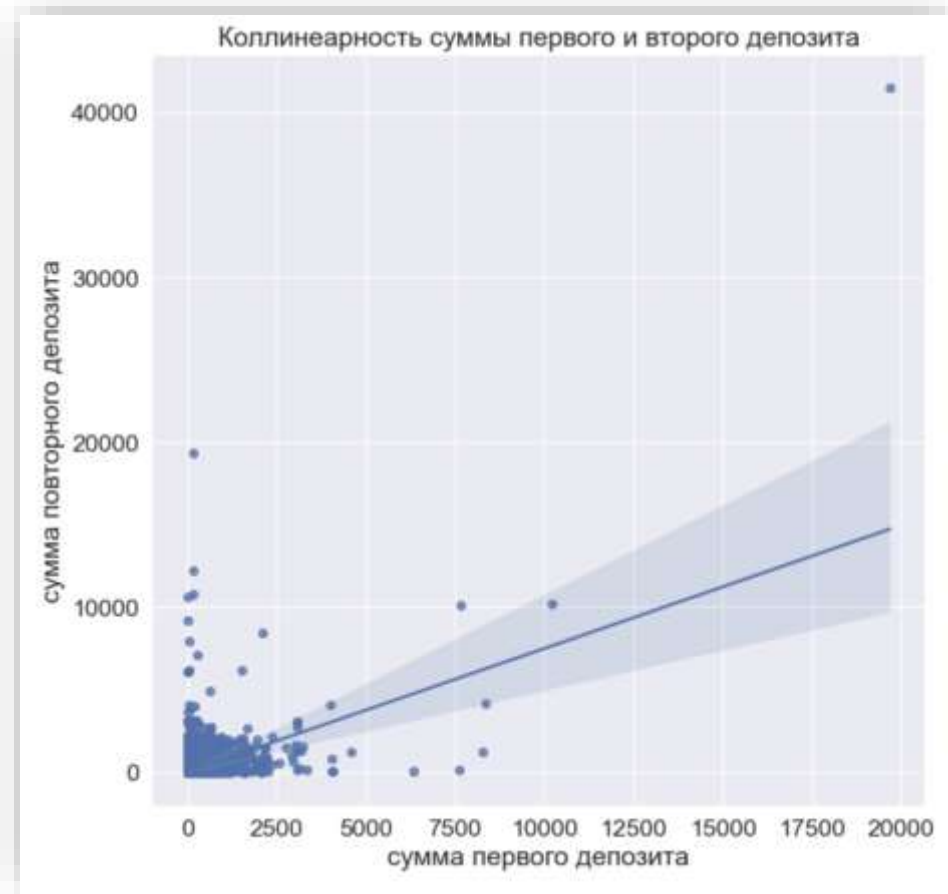
Проверка коллинеарности



Проверка коллинеарности с помощью графика `seaborn.regplot()`

Проверка коллинеарности

Из-за наличия коллинеарности не результирующих показателей было принято решение оставить один столбец из трех - “Сумма инвестиций”, остальные удалить



Повторный корреляционный анализ данных

Тепловая карта корреляции показателей после удаления коллинеарных столбцов

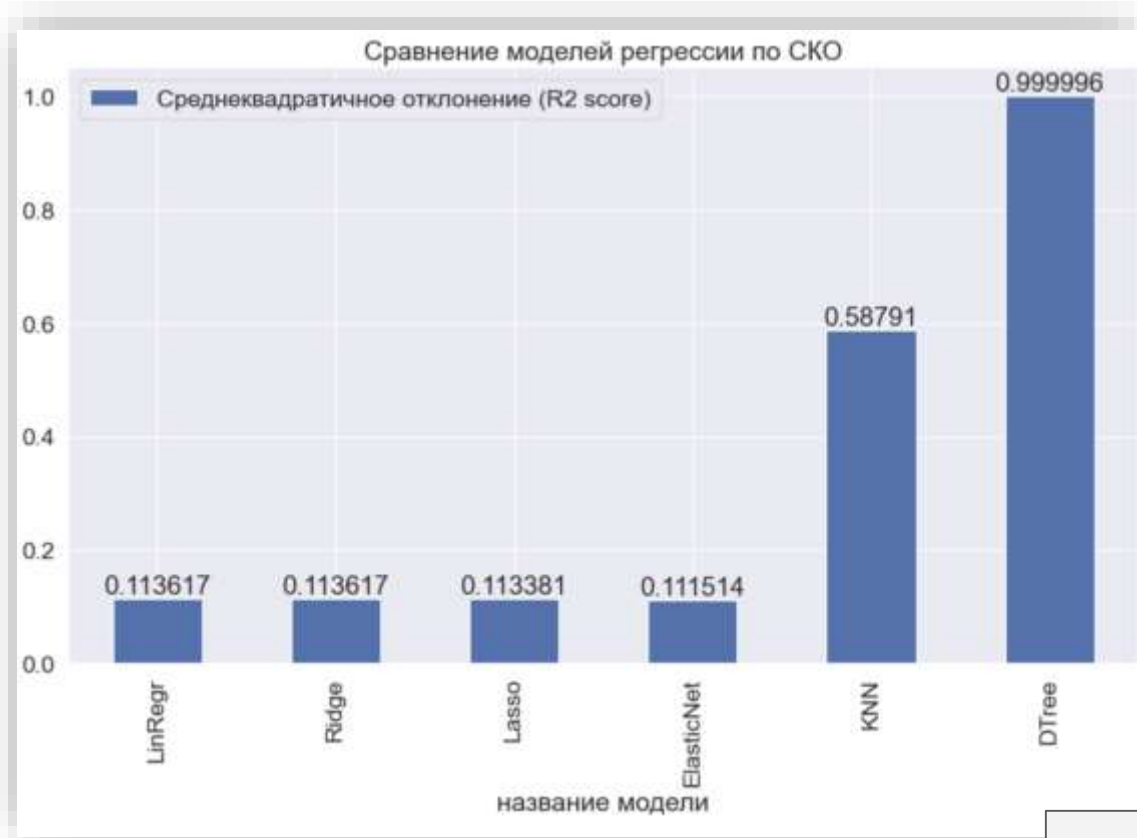


– Тепловая карта корреляции показала наибольшую корреляцию показателя прибыли с суммой инвестиций и инвестиционным периодом

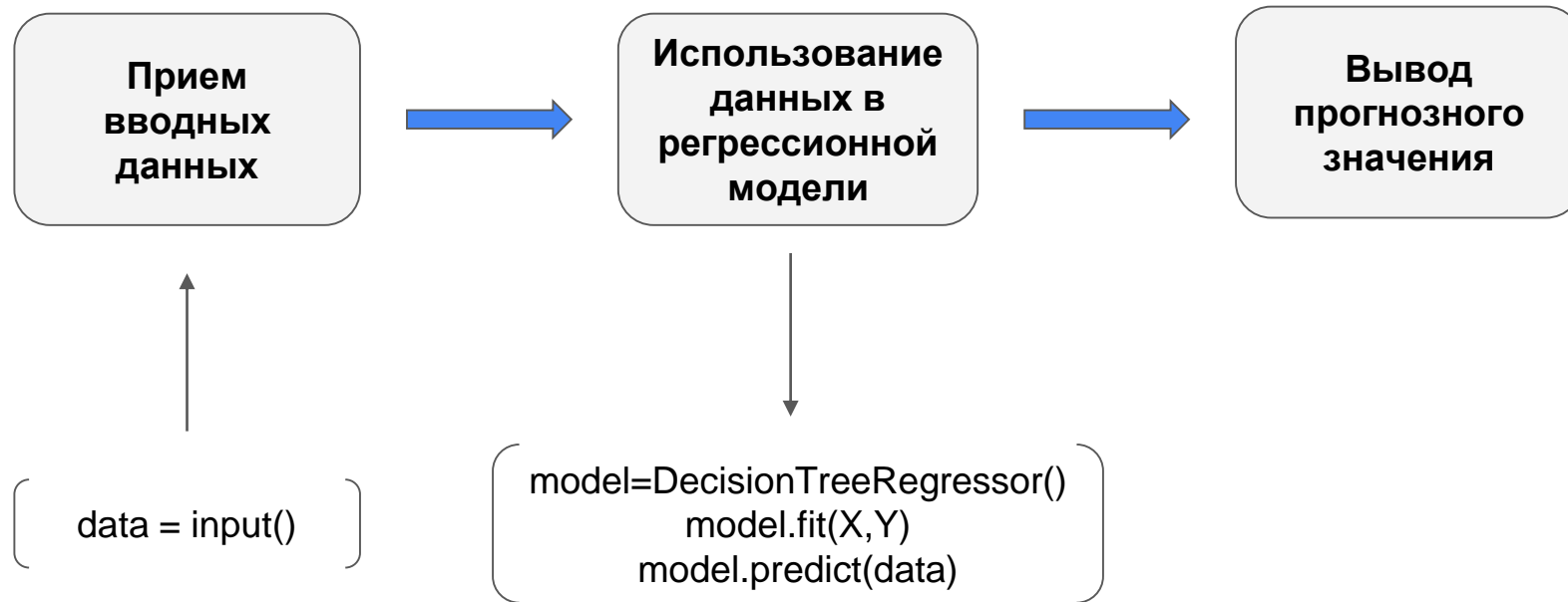
Выбор модели прогнозирования

- модель линейной регрессии (LinRegr)
- модель гребневой регрессии (Ridge)
- модель регрессии Лассо (Lasso)
- модель эластичная сеть (ElasticNet)
- модель ближайшего соседа (KNN)
- модель дерева решений (DTree)

Лучше всех оказалась модель дерева решений (DTree), показав наибольшее СКО = 0.99 и наименьшую среднеквадратичную ошибку MSE = 0.43.



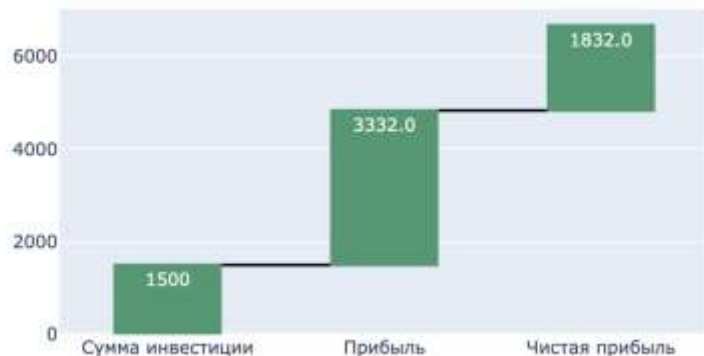
Блок-схема разработанного инструмента



Тестирование разработанного инструмента

	Показатель	Значение
0	Вы выбрали тип актива:	currency
1	Вы выбрали уровень риска портфеля:	high
2	Вы указали размер инвестиции, USD:	1500
3	Вы указали срок инвестиции:	365
4	Вы указали объем оборота активов:	70
5	Ваша прогнозируемая прибыль, USD:	3332.0
6	Чистая прибыль (прибыль - инвестиции)	1832.0

Инфографика прибыли



	Показатель	Значение
0	Вы выбрали тип актива:	index
1	Вы выбрали уровень риска портфеля:	low
2	Вы указали размер инвестиции, USD:	5000
3	Вы указали срок инвестиции:	180
4	Вы указали объем оборота активов:	35
5	Ваша прогнозируемая прибыль, USD:	3581.0
6	Чистая прибыль (прибыль - инвестиции)	-1419.0

Инфографика прибыли



Тестирование разработанного инструмента

При неудачном сценарии, если данных по запрашиваемому типу инвестиционного инструмента нет в базе и прогноз построить невозможно, пользователь получает соответствующее уведомление.

Введите тип актива для инвестиции:

- 1 - currency
- 2 - crypto
- 3 - commodity
- 4 - index stocks

Вы выбрали тип актива: stocks

	Показатель	Значение
0	Вы выбрали тип актива:	stocks
1	Вы выбрали уровень риска портфеля:	ошибка
2	Вы указали размер инвестиции, USD:	ошибка
3	Вы указали срок инвестиции:	ошибка
4	Вы указали объем оборота активов:	ошибка
5	Ваша прогнозируемая прибыль, USD:	ошибка
6	Чистая прибыль (прибыль - инвестиции)	ошибка

Данного инвестиционного инструмента нет в базе! Попробуйте еще раз

Выводы

В результате работы было выполнено:

- Собраны исторические данные, на основе которых работает модель прогнозирования
- Произведен анализ и обработка собранных данных для корректной работы модели
- Выбрана наилучшая модель по величине СКО – “модель дерева решений” (СКО → 1)
- Реализован инструмент для прогноза на основе вводных данных пользователя
- Проведено успешное тестирование инструмента прогнозирования

Выводы

Благодарю за внимание!

