Министерство образования Московской области

Государственное бюджетное профессиональное образовательное учреждение Московской области «Физико-технический колледж»

ЗАЩИЩЕНО

ОЦЕНКА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель специальности 09.02.07 Информационные

системы и программирование

\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_

СИСТЕМА ОПРЕДЕЛЕНИЯ СКЛОННОСТИ КЛИЕНТА К ПРИОБРЕТЕНИЮ МАШИНОМЕСТА

Пояснительная записка к курсовому проекту

**по ПМ.05 Разработка и администрирование систем искусственного интеллекта**

Руководитель КП Разработал

преподаватель студент группы ИСП 4-5

ГБПОУ МО «Физтех-колледж»

\_\_\_\_\_\_\_\_\_\_ Г.В. Базяк \_\_\_\_\_\_\_\_\_\_И.Р. Доля

Долгопрудный, 2024

**Министерство образования Московской области**

**Государственное бюджетное профессиональное образовательное учреждение Московской области «Физико-технический колледж»**

|  |
| --- |
| Специальность 09.02.07  Информационные системы и  Программирование  Квалификация «Программист» |

**УТВЕРЖДЕНО**

цикловой комиссией специальности 09.02.07

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_2024 года

Руководитель специальности

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ЗАДАНИЕ

Для курсового проектирования по МДК 05.02 «Разработка и администрирование систем искусственного интеллекта» студенту 4 курса группы ИСП 4-5

Доли Ивану Романовичу

(фамилия, имя, отчество)

Тема задания: Система определения склонности клиента к приобретению машиноместа\_

I Пояснительная записка

[Error: Reference source not foundError: Reference source not found](#_Toc177754924)Error: Reference source not found

1 Анализ предметной области

1.1 Описание предметной области

1.2 Определение назначения системы и целевой группы пользователей

1.3 Определение ограничений проектного решения

2 Определение требований к системе

2.1 Функциональные требования

2.2 Требования к графическому интерфейсу

2.3 Техническое задание

3 Этапы разработки

3.1 Сбор и анализ данных

3.2 Разработка модели

3.3 Разработка интерфейса

3.4 Тестирование системы

Заключение

Список используемых источников

Приложение

Приложение А.

Приложение Б.

Дата выдачи: « » 2024г. Срок окончания: « » 2024 г

Руководитель курсового проектирования Г.В. Базяк

СОДЕРЖАНИЕ

[Введение 4](#__RefHeading___Toc31153_3180129470)

[1 Анализ предметной области 6](#__RefHeading___Toc31155_3180129470)

[1.1 Описание предметной области 6](#__RefHeading___Toc31157_3180129470)

[1.2 Определение назначения системы и целевой группы пользователей 7](#__RefHeading___Toc31159_3180129470)

[1. 3. Определение ограничений проектного решения 8](#__RefHeading___Toc31161_3180129470_%25D0)

[2 Определение требований к системе 11](#__RefHeading___Toc31163_3180129470)

[2.1 Функциональные требования 11](#__RefHeading___Toc31165_3180129470)

[2.2 Требования к графическому интерфейсу 11](#__RefHeading___Toc31167_3180129470)

[2.3 Техническое задание 12](#__RefHeading___Toc31169_3180129470)

[3 Этапы разработки 12](#__RefHeading___Toc31171_3180129470)

[3.1 Сбор и анализ данных 12](#__RefHeading___Toc31173_3180129470)

[3.2 Разработка модели 13](#__RefHeading___Toc31175_3180129470)

[3.3 Разработка интерфейса 13](#__RefHeading___Toc31177_3180129470)

[3.4 Тестирование системы 14](#__RefHeading___Toc31179_3180129470)

[Заключение 16](#__RefHeading___Toc31181_3180129470)

[Список используемых источников 18](#__RefHeading___Toc31183_3180129470)

[Приложение 19](#__RefHeading___Toc31185_3180129470)

[Приложение А 19](#__RefHeading___Toc31187_3180129470)

[Приложение Б 20](#__RefHeading___Toc31189_3180129470)

# Введение

Современный мир характеризуется быстрым развитием технологий и ростом урбанизации, что приводит к увеличению числа автомобилей и, как следствие, к необходимости оптимизации использования парковочных мест. Для бизнеса, связанного с продажей или сдачей в аренду машиномест, крайне важно эффективно прогнозировать спрос на их услуги. Это требует использования современных аналитических подходов и инструментов, позволяющих оценивать склонность клиентов к приобретению машиномест.

Целью данной курсовой работы является разработка системы определения склонности клиента к приобретению машиноместа на основе анализа данных с использованием методов машинного обучения. Это позволит не только повысить эффективность работы компании, предлагающей парковочные места, но и улучшить качество обслуживания клиентов за счет индивидуального подхода.

Для выполнения задачи предполагается использование существующего набора данных, содержащего информацию о клиентах. Набор данных не подвергался предварительной обработке, что создает дополнительные вызовы, связанные с очисткой, анализом и подготовкой данных для машинного обучения. В рамках работы планируется исследовать основные этапы подготовки данных, выбрать подходящие алгоритмы машинного обучения и оценить их качество. Особое внимание будет уделено интерпретируемости полученной модели, чтобы бизнес мог использовать ее выводы для принятия обоснованных решений.

Актуальность исследования обусловлена возрастающей ролью анализа данных и технологий машинного обучения в управлении бизнес-процессами. Разработка подобной системы позволит не только оптимизировать использование парковочных площадей, но и повысить экономическую эффективность организации.

# 1 Анализ предметной области

## 1.1 Описание предметной области

Управление парковочным пространством в городских условиях — одна из ключевых задач, возникающих на пересечении урбанистики и коммерции. С ростом плотности населения и количества автомобилей необходимость оптимального использования парковочных площадей становится все более актуальной. На практике это включает обеспечение клиентов подходящими условиями для размещения их транспортных средств, что требует учета множества факторов: местоположения, доступности, стоимости и индивидуальных предпочтений клиентов.

Приобретение машиномест часто связано с решением долгосрочных задач пользователей, таких как обеспечение удобного доступа к своему автомобилю или защита от неблагоприятных внешних факторов. В данном случае в качестве клиентов рассматриваются частные лица, владельцы автомобилей, которые принимают решение о покупке машиноместа на основе индивидуальных потребностей, таких как удобство, финансовая доступность, близость к месту жительства или работы.

Одним из ключевых аспектов предметной области является динамика спроса на машиноместа, зависящая от таких факторов, как уровень дохода клиента, расстояние от места проживания или работы до парковочной зоны, сезонные колебания и наличие альтернативных решений (например, использование общественного транспорта или аренда временного парковочного пространства). Понимание этих факторов играет решающую роль при разработке системы анализа и прогнозирования поведения клиентов.

Таким образом, предметная область находится на стыке анализа данных, маркетинга и управления недвижимостью. Ее изучение позволяет разработать эффективные инструменты, способные учитывать широкий спектр характеристик клиентов и факторов, влияющих на их склонность к покупке машиноместа.

## 1.2 Определение назначения системы и целевой группы пользователей

Назначение системы заключается в определении склонности клиентов фирмы, например, строительной компании или застройщика, к приобретению машиноместа. Система будет служить инструментом для анализа данных о потенциальных покупателях и предоставления прогноза, указывающего, с какой вероятностью конкретный клиент совершит покупку. Это позволит компании оптимизировать процесс работы с клиентами, сконцентрировав усилия на тех, кто наиболее вероятно заинтересован в приобретении машиноместа.

Целевая группа пользователей системы включает сотрудников компании, ответственных за продажи и управление клиентскими базами данных. Среди них:

* Менеджеры по продажам, которые используют прогнозы системы для приоритизации работы с клиентами. На основе предоставленных предположений менеджеры могут сосредотачиваться на взаимодействии с клиентами, имеющими высокую вероятность покупки.
* Руководители отделов продаж, которые могут использовать агрегированные результаты работы системы для планирования ресурсов и повышения общей эффективности процесса продаж.
* Аналитики, занимающиеся оценкой эффективности продаж и анализа клиентской базы. Система предоставляет им дополнительные данные для проверки гипотез и понимания факторов, влияющих на успешность сделок.

Система предназначена исключительно для прогнозирования вероятности покупки машиноместа конкретным клиентом. Она фокусируется на индивидуальном уровне принятия решения.

Таким образом, ключевое назначение системы заключается в упрощении принятия решений для сотрудников компании, связанных с продажами, и в повышении эффективности взаимодействия с клиентами на основе данных.

## 1. 3. Определение ограничений проектного решения

Разработка системы определения склонности клиента к приобретению машиноместа предполагает наличие ряда ограничений, связанных как с особенностями предметной области, так и с применением технологий машинного обучения. Основные ограничения проектного решения включают следующие аспекты:

1. Ограничения на исходные данные

* Качество данных: Так как данные для модели собираются автоматически, их качество напрямую влияет на точность прогнозов. Ошибки при сборе данных, неполные или несбалансированные выборки могут существенно снизить производительность системы.
* Динамичность данных: Изменения в источниках данных или в их структуре могут исказить результаты работы системы. В некоторых случаях это потребует полной перенастройки модели или даже ее повторного обучения на новом наборе данных.
* Недостаток признаков: Если автоматически собираемые данные не включают критически важные признаки, влияющие на склонность клиента к покупке (например, данные о доходах или местоположении), это ограничит точность модели.

2. Ограничения модели

* Сложность интерпретации: Алгоритмы машинного обучения, особенно сложные (например, градиентный бустинг или глубокие нейронные сети), могут быть трудны для интерпретации, что затрудняет понимание причин конкретного прогноза.
* Переобучение: При работе с автоматически собранными данными существует риск переобучения модели, что приводит к снижению ее эффективности на новых данных.
* Генерализация: Модель, обученная на данных одной выборки, может быть плохо применима к другим наборам данных, особенно при изменении структуры или содержания исходных данных.

3. Ограничения использования

* Требования к актуальности: Автоматически собираемые данные могут быстро устаревать, что снизит эффективность прогнозов. Для поддержания актуальности модели потребуется регулярное обновление данных и, возможно, повторное обучение.
* Учет человеческого фактора: Система оценивает вероятность покупки на основе данных, но не может учитывать все субъективные и эмоциональные аспекты, влияющие на принятие решений клиентом.
* Ограниченная функциональность: Система предоставляет только вероятностную оценку склонности клиента к покупке и не формирует рекомендации для менеджеров по продажам.

# 2 Определение требований к системе

## 2.1 Функциональные требования

Система определения склонности клиента к приобретению машиноместа должна обеспечивать выполнение следующих функциональных задач:

* Принимать собранные данные о клиентах в установленном формате (файл формата csv)
* Проводить предварительную обработку данных, включая обработку пропусков и нормализацию числовых данных.
* На основе входных данных возвращать бинарный результат: "купит" или "не купит" ли клиент машиноместо.
* Сохранять результаты прогнозов для последующего анализа.

## 2.2 Требования к графическому интерфейсу

Графический интерфейс системы должен быть интуитивно понятным и удобным для пользователя, выполнять следующие функции:

* позволять загрузить данные
* обработать их
* предварительно просмотреть результат
* сохранить результат

## 2.3 Техническое задание

Цели:

Создать систему, которая позволит предсказать склонен ли клиент к покупке машиноместа.

Основные функции системы: создание прогноза на основе автоматически собранных данных. Интерфейс для загрузки и обработки данных, предпросмотра и сохранения результата.

Платформа и инструменты:

Язык программирования: Python.

Библиотеки для обработки данных и машинного обучения: pandas, NumPy, CatBoostClassifier, pickle.

Библиотека для создания интерфейса: tkinter

Архитектура системы: монолитная, десктопное приложение.

# 3 Этапы разработки

## 3.1 Сбор и анализ данных

В данном случае сбор данных не требуется, т. к. используются уже готовые датасеты предоставленные компанией «Самолет». Большая часть данных представляет собой просто пронумерованные столбцы, т. е. существуют множества значений, которые нет возможности идентифицировать, поэтому понять какие факторы повлияют на прогноз — практически невозможно.

В ходе анализа данных было решено удалить столбцы, где пропущено больше 50% Также в ходе анализа был удален столбец содержащий неизвестные хеши ввиду невозможности определить что в них зашифровано, а следовательно невозможности правильно перекодировать данные для обучения модели. Далее все пропуски были заполнены нулевыми значениями. На этом предобработка данных была завершена.

## 3.2 Разработка модели

Для решения поставленной задачи был выбран алгоритм градиентного бустинга CatBoostClassifier. Данный выбор обусловлен следующими причинами:

* Высокая точность прогнозов: CatBoostClassifier демонстрирует высокую точность прогнозов благодаря использованию градиентного бустинга и различных методов регуляризации.
* Скорость обучения и масштабируемость: Алгоритм оптимизирован для работы с большими объемами данных и позволяет строить модели в разумные сроки.
* Удобство использования: Простой интерфейс библиотеки CatBoost и гибкие настройки гиперпараметров делают алгоритм удобным в использовании.

## 3.3 Разработка интерфейса

Интерфейс на tkinter устроен следующим образом:

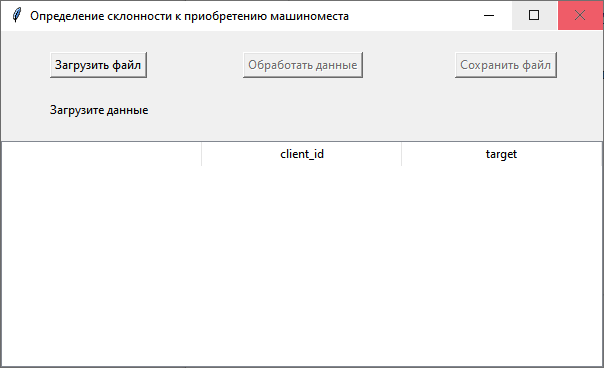
Сверху в ряд располагаются три кнопки: "Загрузить файл", "Обработать данные"и "Сохранить файл".

Ниже, в ряд располагается три надписи (объект label), в них отображается статус выполнения программы.

В самом низу располагается таблица, для предпросмотра результатов.

В начале работы с приложением активна только одна кнопка: "Загрузить файл". Загрузка выполняется посредством диалогового окна выбора файла. После загрузки становится доступна кнопка "Обработать данные", а после обработки соответственной и "Сохранить файл". Сохранение выполняется посредством соответствующего диалогового окна. После обработки данных, они отображаются в таблице. В надписи под кнопкой загрузки изначально содержится текст "Загрузите данные", после загрузки он меняется на "Данные загружены!". После обработки под кнопкой обработки появляется текст "Данные обработаны!", а под кнопкой сохранения "Сохранить данные?". После сохранения она меняется на "Данные сохранены!". При повторной загрузке данных последние две надписи очищаются.

Таким образом получается минималистичный и интуитивно понятный интерфейс.

Рис. 1 Пользовательский интерфейс.

## 3.4 Тестирование системы

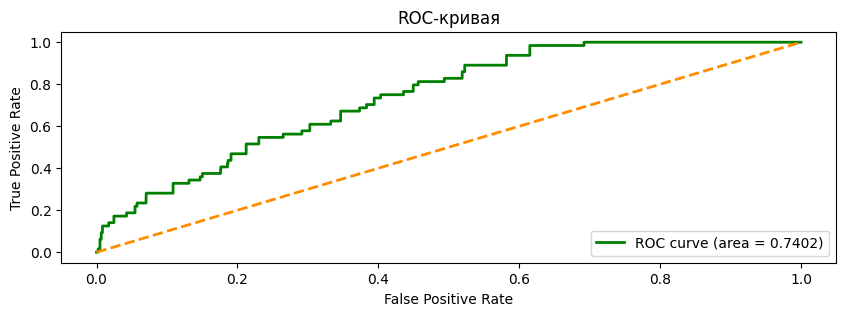
Объяснение результатов:

ROC-кривая (Receiver Operating Characteristic) – это графический инструмент, который позволяет оценить качество бинарного классификатора. Она отображает соотношение между долей истинно положительных классификаций (True Positive Rate, TPR) и долей ложно положительных классификаций (False Positive Rate, FPR) при различных порогах классификации.

Порог для "хорошего" значения AUC зависит от конкретной задачи и допустимого уровня ошибок. Однако, в общем случае:

* AUC < 0.7: Неудовлетворительный результат, модель работает плохо и требует существенной доработки.
* 0.7 < AUC < 0.8: Удовлетворительный результат, модель может быть использована, но возможно потребуется дополнительная настройка.
* AUC > 0.8: Считается хорошим результатом, модель демонстрирует высокую способность различать классы.
* AUC > 0.95: Может считаться отличным результатом, хотя с большей вероятностью свидетельствует о переобучении модели.

В нашем случае ROC-кривая показала 0.74, что говорит об удовлетворительном результате.

Рис. 2 ROC-Кривая модели

# Заключение

В рамках данной курсовой работы была разработана и внедрена система прогнозирования склонности клиентов к приобретению машиноместа основанная на использовании методов машинного обучения. Система ориентирована на обработку данных, предоставленных компанией, и позволяет предсказать вероятность покупки машиноместа для каждого клиента на основе множества факторов.

Для построения модели был выбран алгоритм CatBoostClassifier, который несмотря на не очень высокое качество данных продемонстрировал отличные результаты при работе с ними и способен эффективно обрабатывать как малые, так и большие объемы данных. Алгоритм продемонстрировал хорошую точность и устойчивость к переобучению, что является важным аспектом в задачах классификации, особенно когда данные далеки от идеала.

На основе собранных данных был произведен анализ и предварительная обработка, в ходе которой были устранены ошибки и выбраны важнейшие признаки, которые могли бы повлиять на результат модели. Важно, работа с данными была организована таким образом, чтобы повысить стабильность и надежность прогнозов, а также уменьшить риски ошибок, связанных с неправильной обработкой данных.

Процесс тестирования системы показал ее способность точно и быстро выполнять прогнозы, что подтверждается хорошими значениями показателей точности, таких как AUC-ROC, и стабильностью результатов на разных выборках. Несмотря на это, система имеет потенциал для дальнейшего развиватия и требует дальнейшего тестирования и оптимизации на новых данных для улучшения точности и адаптации к возможным изменениям в поведении клиентов.

В будущем следует рассмотреть возможность реализации автоматического анализа данных по мере их попадания в базу данных компании, а также автоматическое дообучение на свежих данных. Это позволит еще сильнее повысить её эффективность и точность.

В результате работы была создана система, которая может быть успешно использована для анализа клиентских данных и предсказания их склонности к приобретению машиномест. Это открывает возможности для повышения эффективности маркетинговых кампаний и продаж, а также для улучшения взаимодействия с клиентами компании, что в свою очередь может привести к росту бизнес-показателей и конкурентоспособности фирмы на рынке недвижимости.

# Список используемых источников

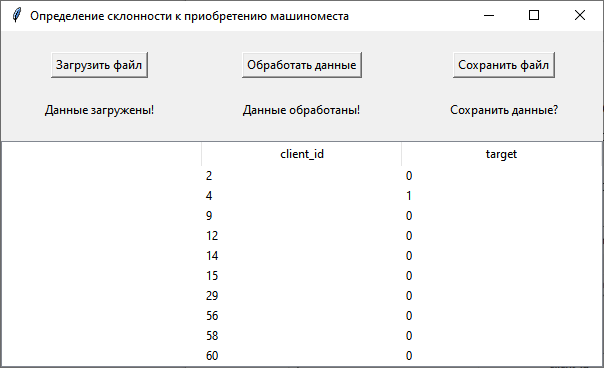
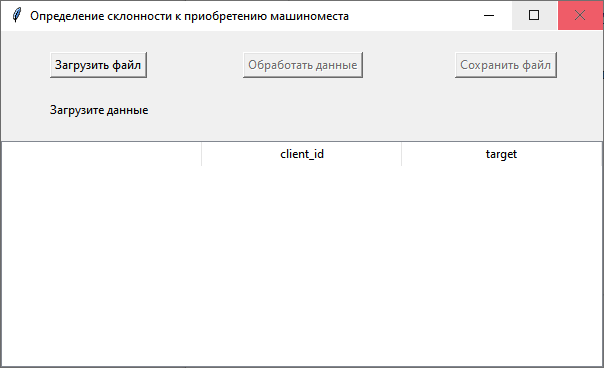
1. Официальная документация Python [Электронный ресурс]. URL: <https://docs.python.org/3/> (Дата обращения 12.12.2024).
2. W3Schools Python Tutorial [Электронный ресурс]. URL: <https://www.w3schools.com/python/> (Дата обращения 12.12.2024).
3. Официальная документация Pandas [Электронный ресурс]. URL: <https://pandas.pydata.org/pandas-docs/stable/> (Дата обращения 13.12.2024).
4. TutorialsPoint: Pandas Tutorial [Электронный ресурс]. URL: <https://www.tutorialspoint.com/python_pandas/index.htm> (Дата обращения 13.12.2024).
5. Официальная документация Python по модулю Pickle [Электронный ресурс]. URL: <https://docs.python.org/3/library/pickle.html> (Дата обращения 15.14.2024).
6. Официальная документация Tkinter [Электронный ресурс]. URL: <https://docs.python.org/3/library/tkinter.html> (Дата обращения 15.14.2024).
7. Руководство по программированию на Tkinter и Python [Электронный ресурс]. URL: <https://metanit.com/python/tkinter/> (Дата обращения 14.12.2024).
8. Курс по машинному обучению от Google (TensorFlow) [Электронный ресурс]. URL: <https://www.tensorflow.org/tutorials> (Дата обращения 15.12.2024).
9. Официальная документация библиотеки CatBoost [Электронный ресурс]. URL: <https://catboost.ai/docs/> (Дата обращения 15.12.2024).

# Приложение

## Приложение А

(обязательное)

Интерфейс

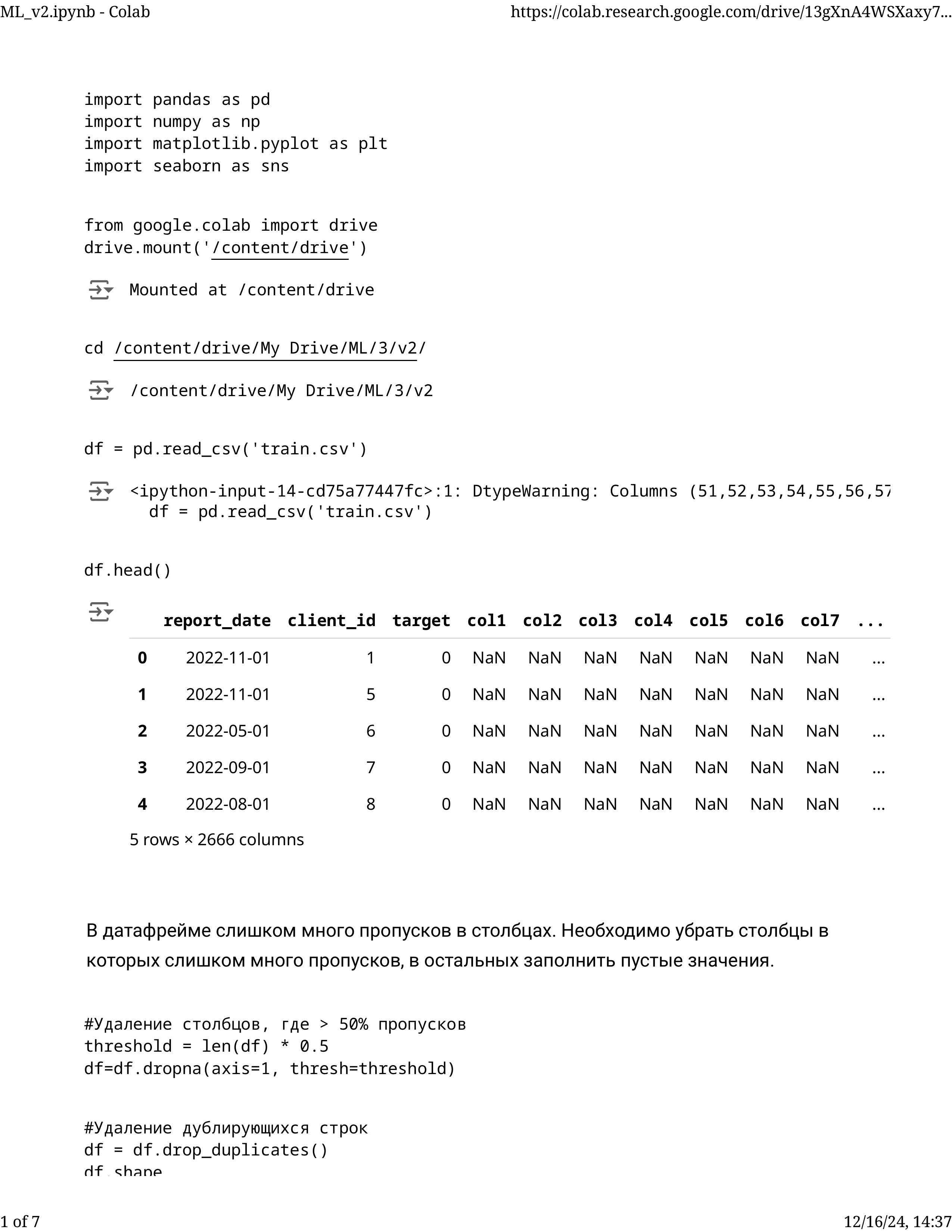


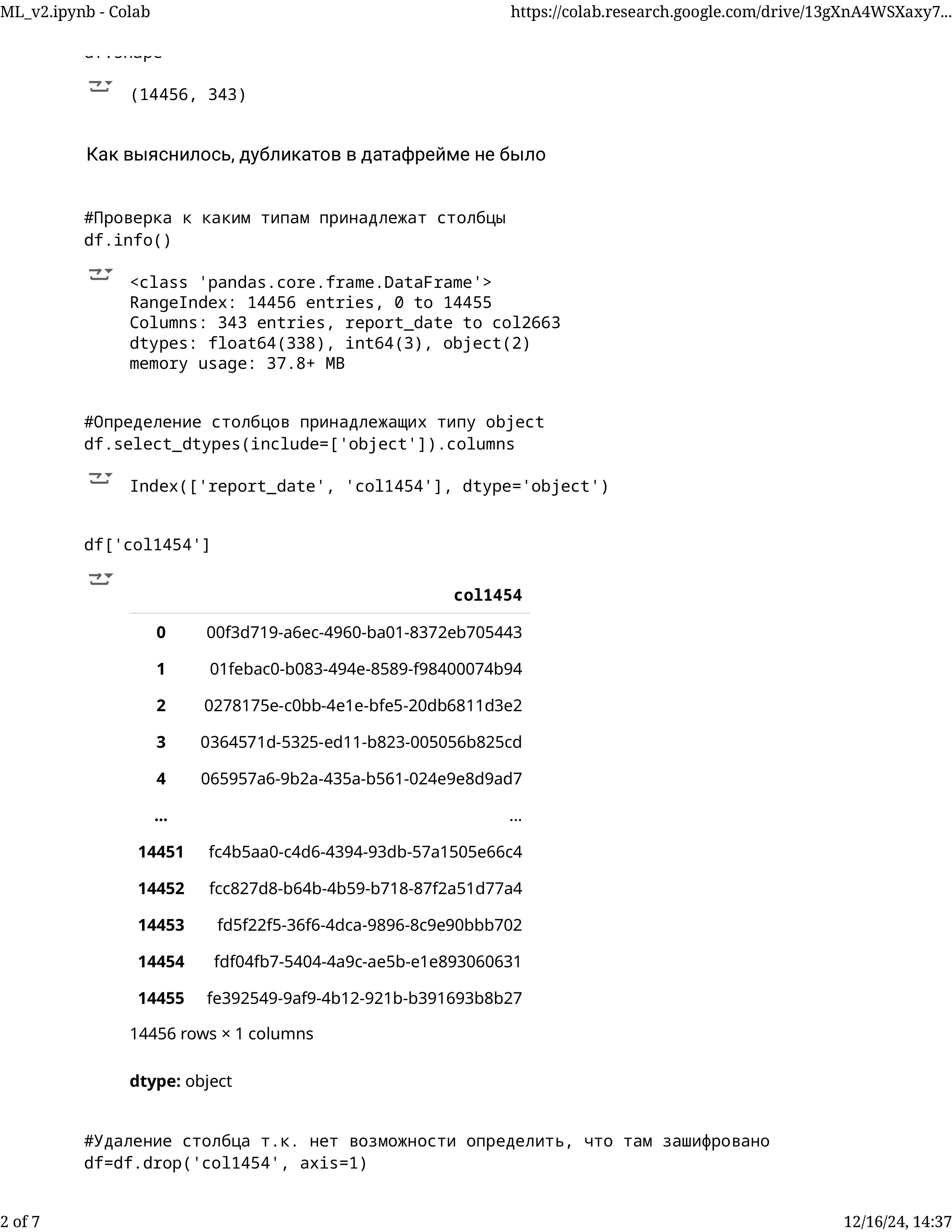
## Приложение Б

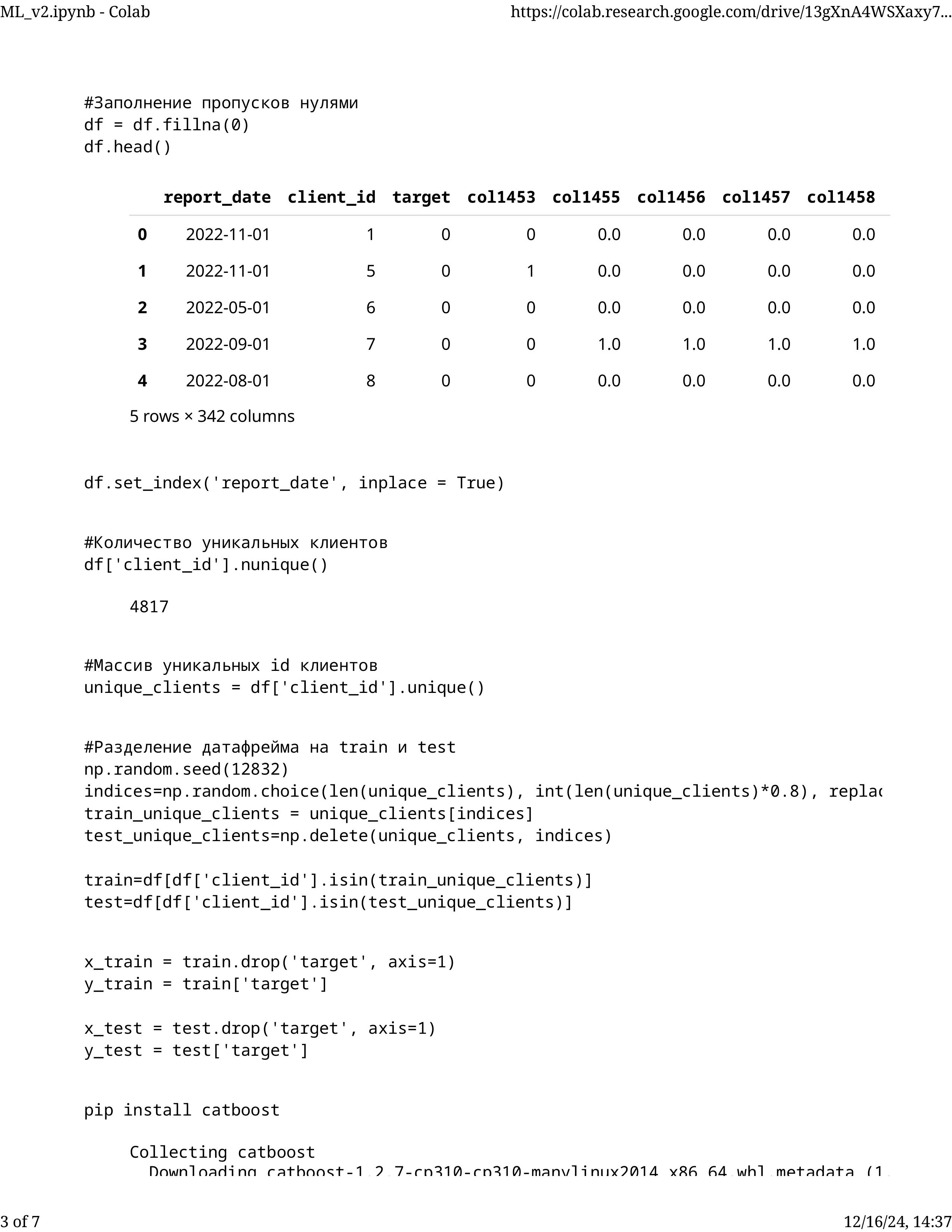
(обязательное)

Листинг программы

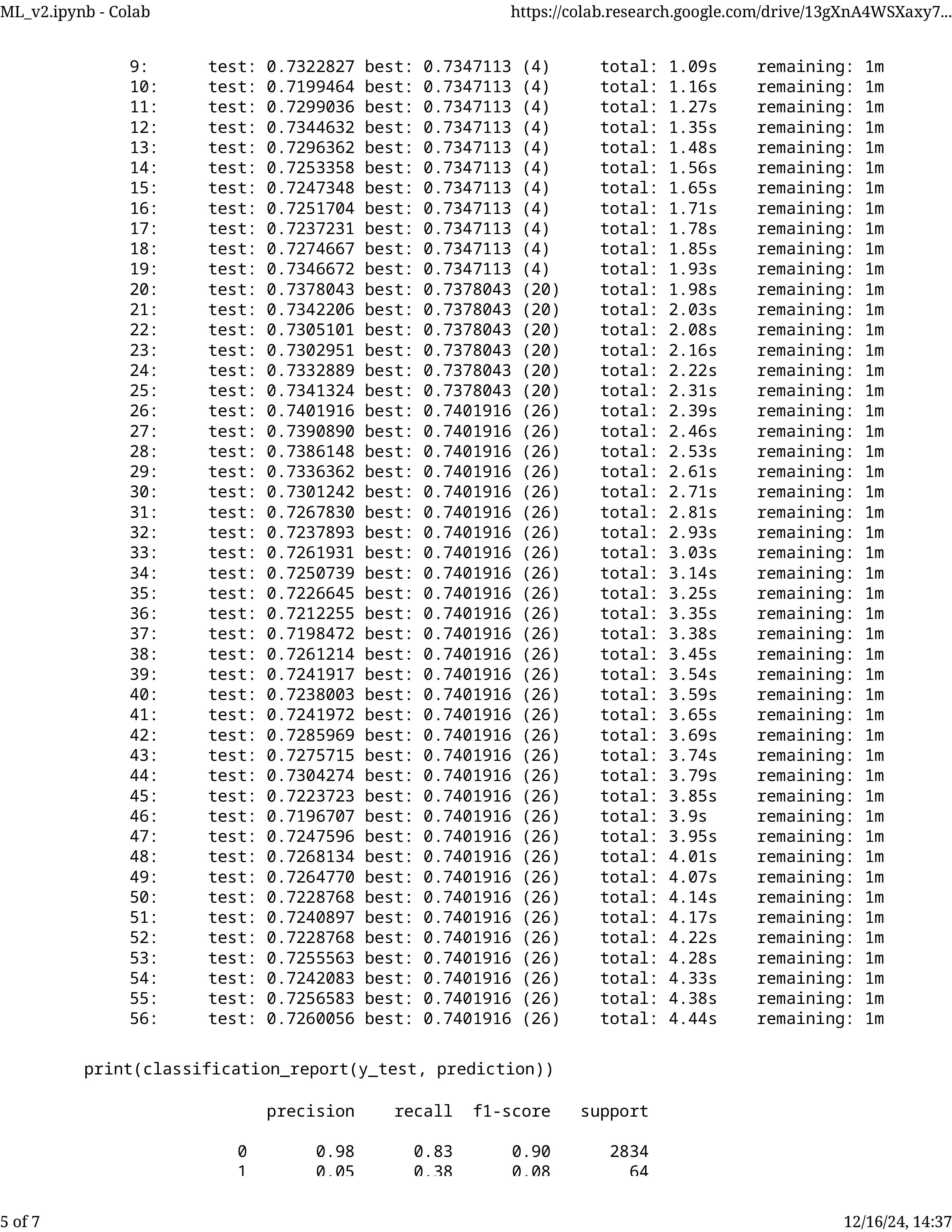
Обучение модели

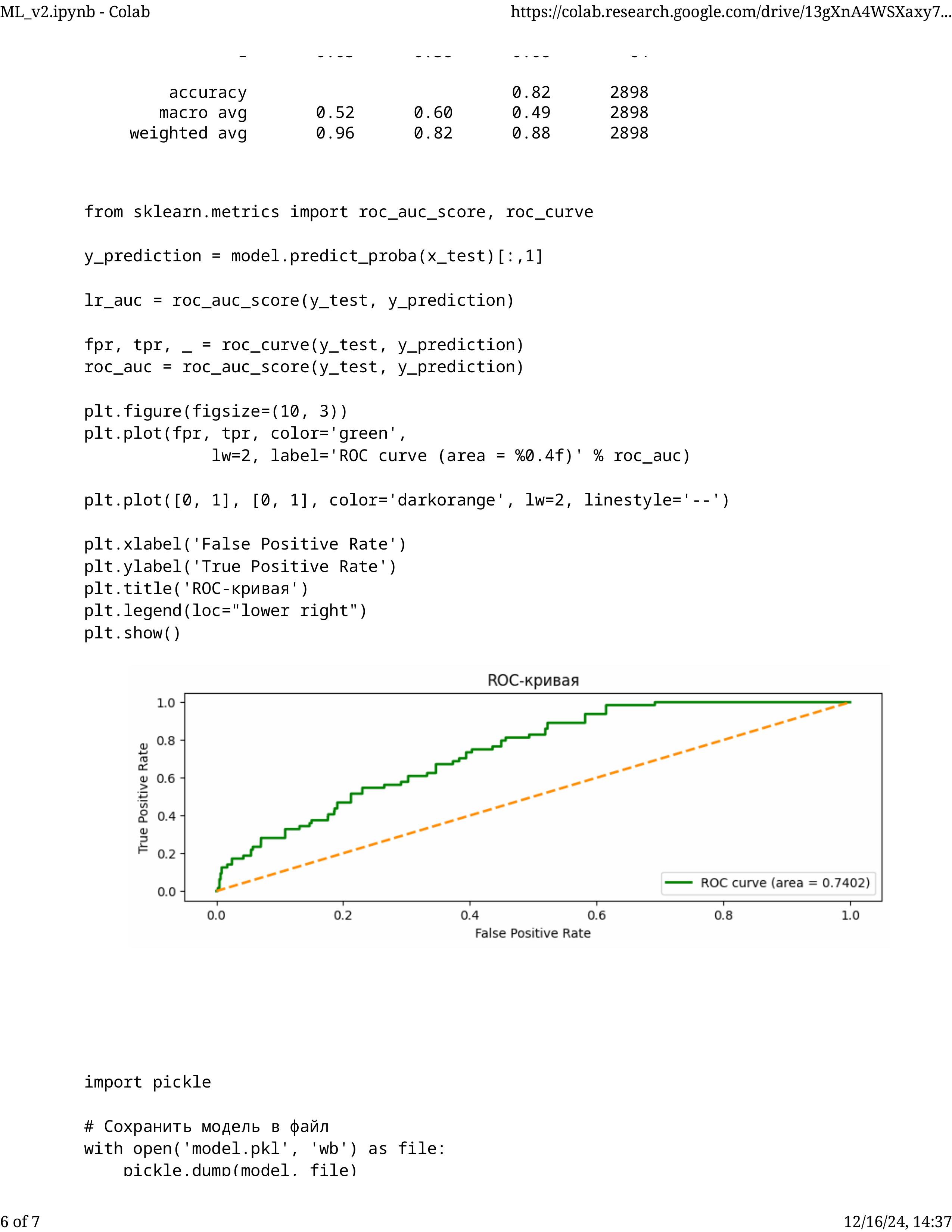












Код приложения

import tkinter as tk

from tkinter import filedialog

from tkinter import ttk

import pandas as pd

import numpy as np

import pickle

from catboost import CatBoostClassifier, Pool

def load\_data():

global input

file\_path = filedialog.askopenfilename(

title="Выберите файл",

filetypes=(("Таблицы", "\*.csv"), ("Все файлы", "\*.\*"))

)

if file\_path: # Если файл выбран

input = pd.read\_csv(file\_path)

table.delete(\*table.get\_children())

process\_button.config(state=tk.NORMAL)

lbl4.config(text="Данные загружены!")

lbl5.config(text="")

lbl6.config(text="")

save\_button.config(state=tk.DISABLED)

def process\_data():

global results

# предобработка

input\_ = input.fillna(0)

input\_.set\_index('report\_date', inplace=True)

# Удаление столбцов не подходящих модели

columns\_to\_keep = ['report\_date'] # Исключаемый столбец

columns\_to\_drop = [

col for col in input\_.select\_dtypes(include='object').columns

if col not in columns\_to\_keep

]

input\_ = input\_.drop(columns=columns\_to\_drop)

# Анализ данных

predictions = model.predict(input\_)

# Результат

client\_ids = input['client\_id']

results = pd.DataFrame({

'client\_id': client\_ids,

'target': predictions

})

lbl5.config(text="Данные обработаны!")

lbl6.config(text="Сохранить данные?")

# Вывод в таблицу

for row in table.get\_children():

table.delete(row)

for \_, row in results.iterrows():

table.insert("", "end", values=(row['client\_id'], row['target']))

save\_button.config(state=tk.NORMAL)

def save\_data():

file\_path = filedialog.asksaveasfilename(defaultextension=".csv")

if file\_path:

results.to\_csv(file\_path, index=False)

lbl6.config(text="Данные сохранены!")

with open('model.pkl', 'rb') as file:

model = pickle.load(file)

# Создаем главное окно

root = tk.Tk()

root.title("Определение склонности к приобретению машиноместа")

# Создание отступов

lbl1 = tk.Label(root)

lbl1.grid(row=4)

lbl2 = tk.Label(root)

lbl2.grid(row=0)

lbl3 = tk.Label(root)

lbl3.grid(row=2)

# Кнопки для загрузки, сохранения и обработки

load\_button = tk.Button(root, text="Загрузить файл", command=load\_data)

load\_button.grid(row=1, column=0)

process\_button = tk.Button(root, text="Обработать данные", command=process\_data, state=tk.DISABLED)

process\_button.grid(row=1, column=1)

save\_button = tk.Button(root, text="Сохранить файл", command=save\_data, state=tk.DISABLED)

save\_button.grid(row=1, column=2)

# Создание полей для вывода статуса

lbl4 = tk.Label(root, text='Загрузите данные')

lbl4.grid(row=3, column=0)

lbl5 = tk.Label(root, text='')

lbl5.grid(row=3, column=1)

lbl6 = tk.Label(root, text='')

lbl6.grid(row=3, column=2)

# Создание таблицы

table = ttk.Treeview(root, columns=("Column1", "Column2"))

table.heading("Column1", text="client\_id")

table.heading("Column2", text="target")

table.grid(row=5, column=0, columnspan=3)

root.mainloop()