|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра ИУ5 «Системы обработки информации и управления»**

Отчет по курсовой работе

Сравнительный анализ GradientBoosting и CatBoost на примере датасета Human Resources

по дисциплине «Машинное обучение»

Выполнил:

Кондратьев Максим

05.06.2019

Проверил:

к.т.н., доц., Ю. Е. Гапанюк

подпись, дата

Москва, 2019 г.

Задача:

На основе датасета Human Resources https://www.kaggle.com/rhuebner/human-resources-data-set провести сравнительный анализ двух схем ансамблевых моделей CatBoost и scikit learn GradientBoosting для решения задачи регрессии. Провести разведочный анализ данных, выбор необходимых признаков для каждой из моделей, корреляционный анализ. Выбрать необходимые метрики, построить базовое решение, провести подбор гиперпараметров моделей. Сформировать вывод о качестве построенных моделей на основе выбранных метрик.

Содержание:

1. Подготовительный этап. Чтение и анализ данных
2. Отбор параметров датасета. Кодирование категориальных параметров. Исследование и обработка данных.
3. Построение модели CatBoostRegressor. Подбор параметров. Построение кривой обучения и валидации на основе лучшей модели.
4. Демонстрация результатов модели CatBoostRegressor.
5. Исследование и обработка данных.
6. Построение модели GradientBoostingRegressor. Подбор параметров. Построение кривой обучения и валидации на основе лучшей модели.
7. Демонстрация результатов модели GradientBoostingRegressor.
8. Вывод. Сравнение двух моделей на основе выбранных метрик. Сравнение моделей по времени обучения и удобству использования.

Введение:

CatBoost — открытая программная библиотека разработанная компанией Яндекс и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга. Основное API для работы с библиотекой реализовано для языка Python, также существует реализация для языка программирования R.

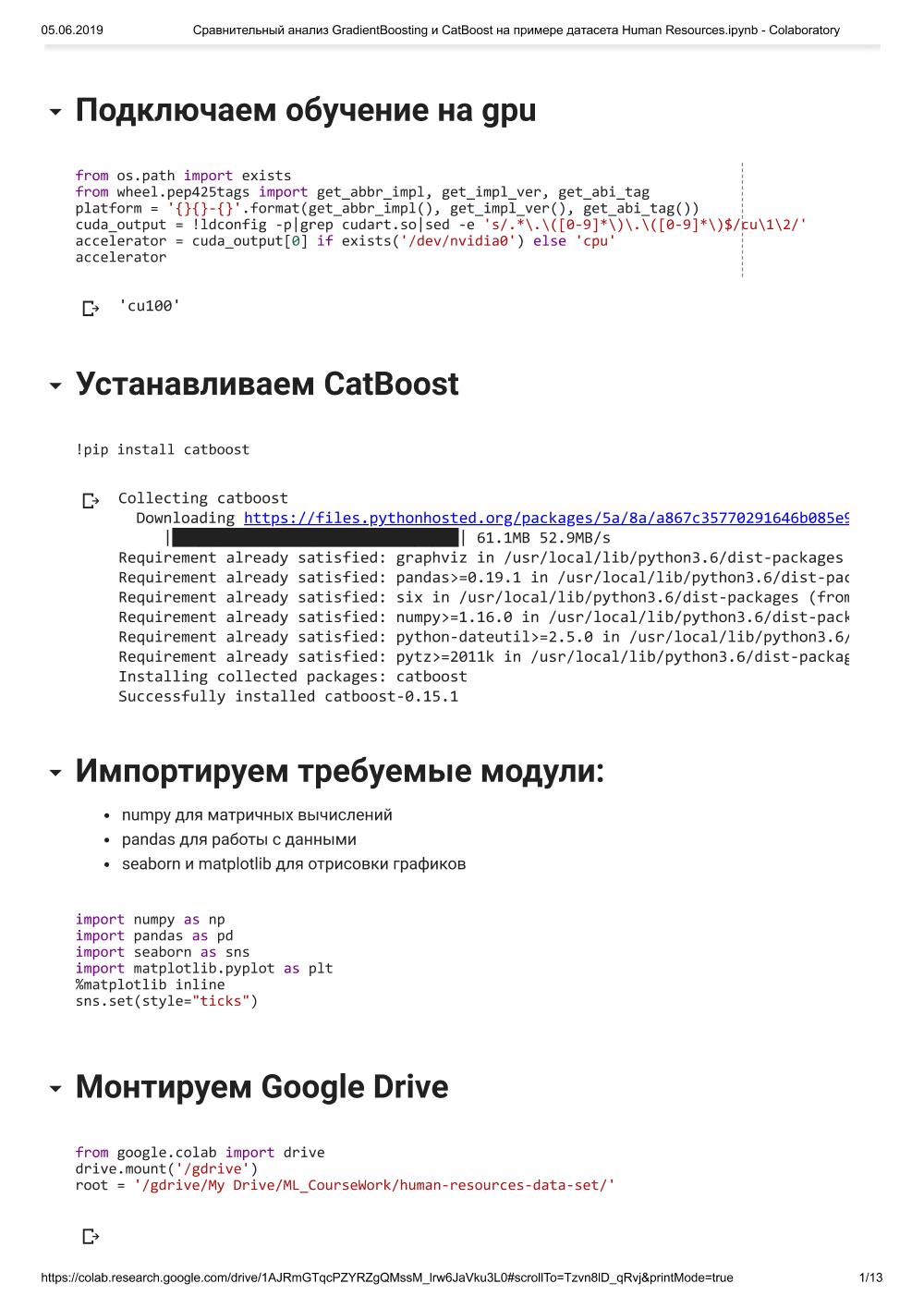
18 июля 2017 года CatBoost была открыта для свободного доступа на GitHub компанией Яндекс под свободной лицензией Apache 2.0. CatBoost является системой машинного обучения, использующая одну из оригинальных схем градиентного бустинга. Сравнивая CatBoost с подобными системами машинного обучения компаний Google (TensorFlow) и Microsoft (LightGBM), руководитель разработки систем машинного обучения «Яндекса» Анна Вероника Дорогуш отметила, что Google TensorFlow решает другой класс задач, эффективно анализируя однородные данные — например изображения. А *«CatBoost работает с данными разной природы и может быть использован в связке с TensorFlow и другими алгоритмами машинного обучения в зависимости от конкретных задач»*. У Microsoft LightGBM российская разработка выигрывает по качеству, что демонстрирует таблица тестов с общепринятыми в машинном обучении сравнениями, но пока проигрывает в скорости — что Яндекс обещает исправить.

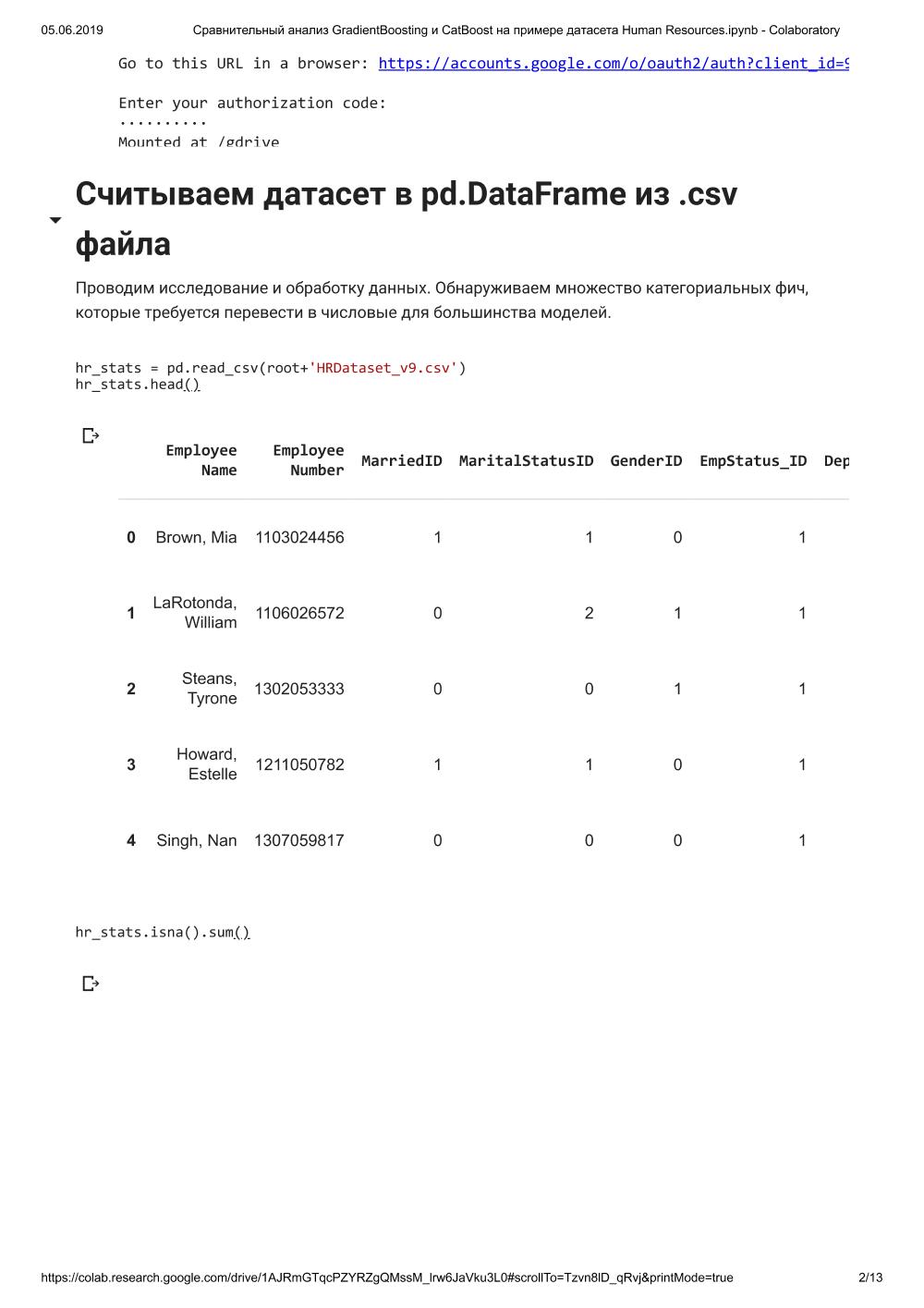
В качестве датасета для исследования данной технологии я решил взять набор данных, в котором наблюдается преобладание категориальных параметров над числовыми. В качестве «эталона» я решил сравнить показатели модели CatBoost со стандартным градиентным бустингом модуля scikit learn.

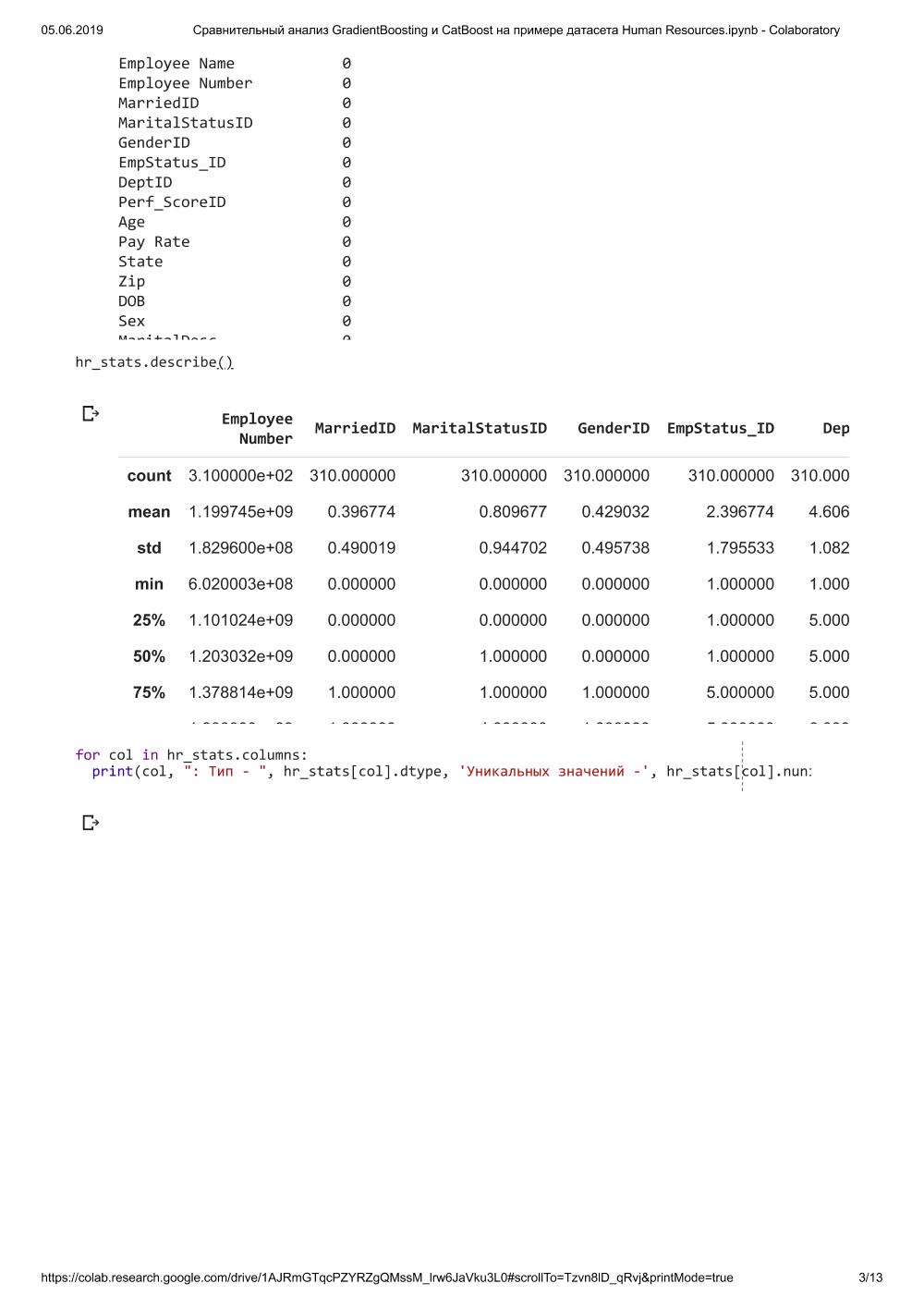
Основная часть:

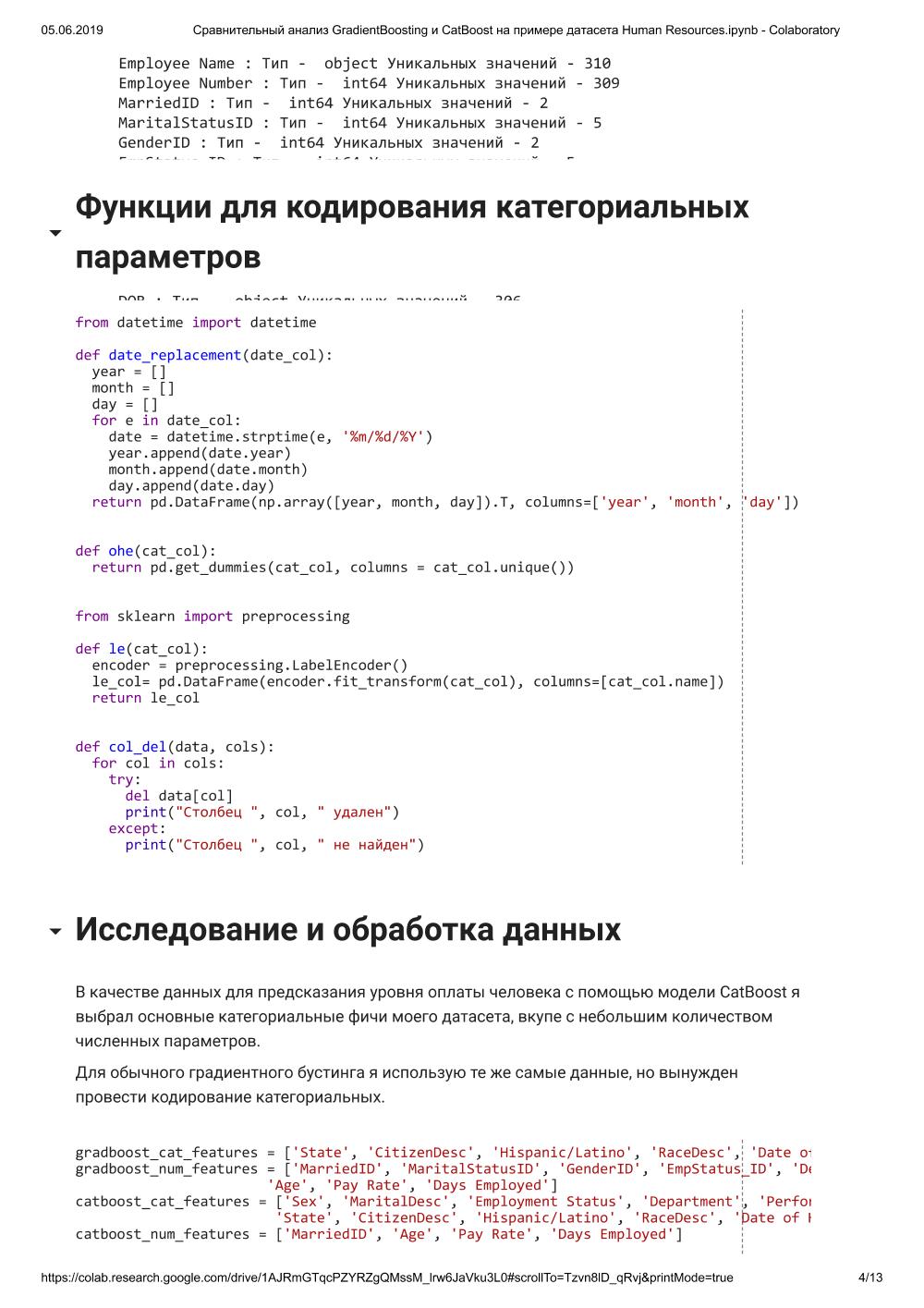
Целью курсовой работы является создание эффективной модели, предсказывающей уровень оплаты работника на основе его характеристик (пола, возраста, эффективности работы, региона проживания, расы, должности, возраста и т. д.). Для решения данной задачи регрессии необходимо провести разведочный анализ данных, отобрать необходимые признаки, провести их обработку для каждой из моделей. Также необходимо выбрать метрики и подобрать параметры для моделей, после чего проанализировать результаты и сделать вывод о качестве построенных моделей.

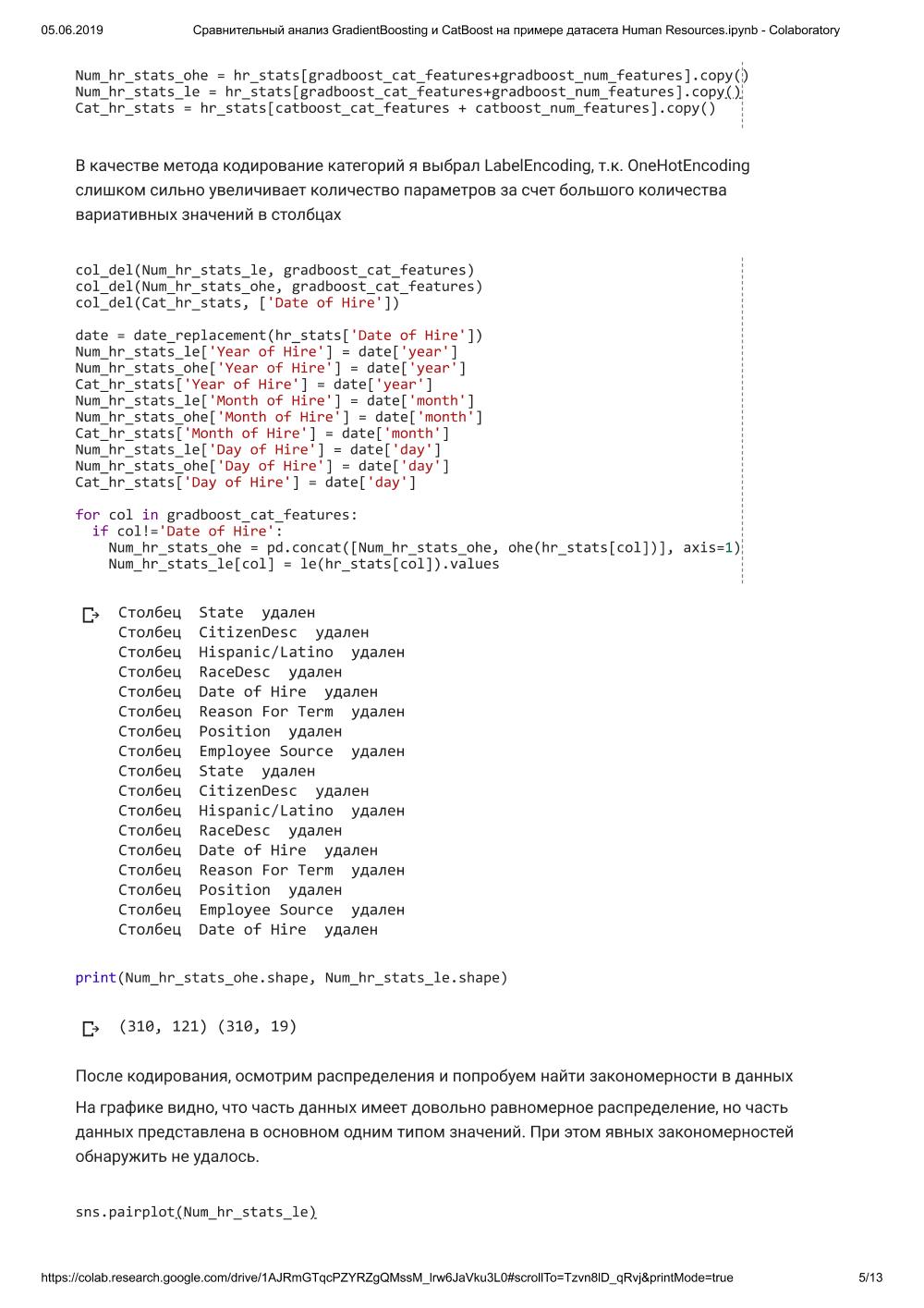
Процесс решения поставленной задачи отображен в виде исходного кода и экранных форм выполнения программы в сервисе Google Colab:

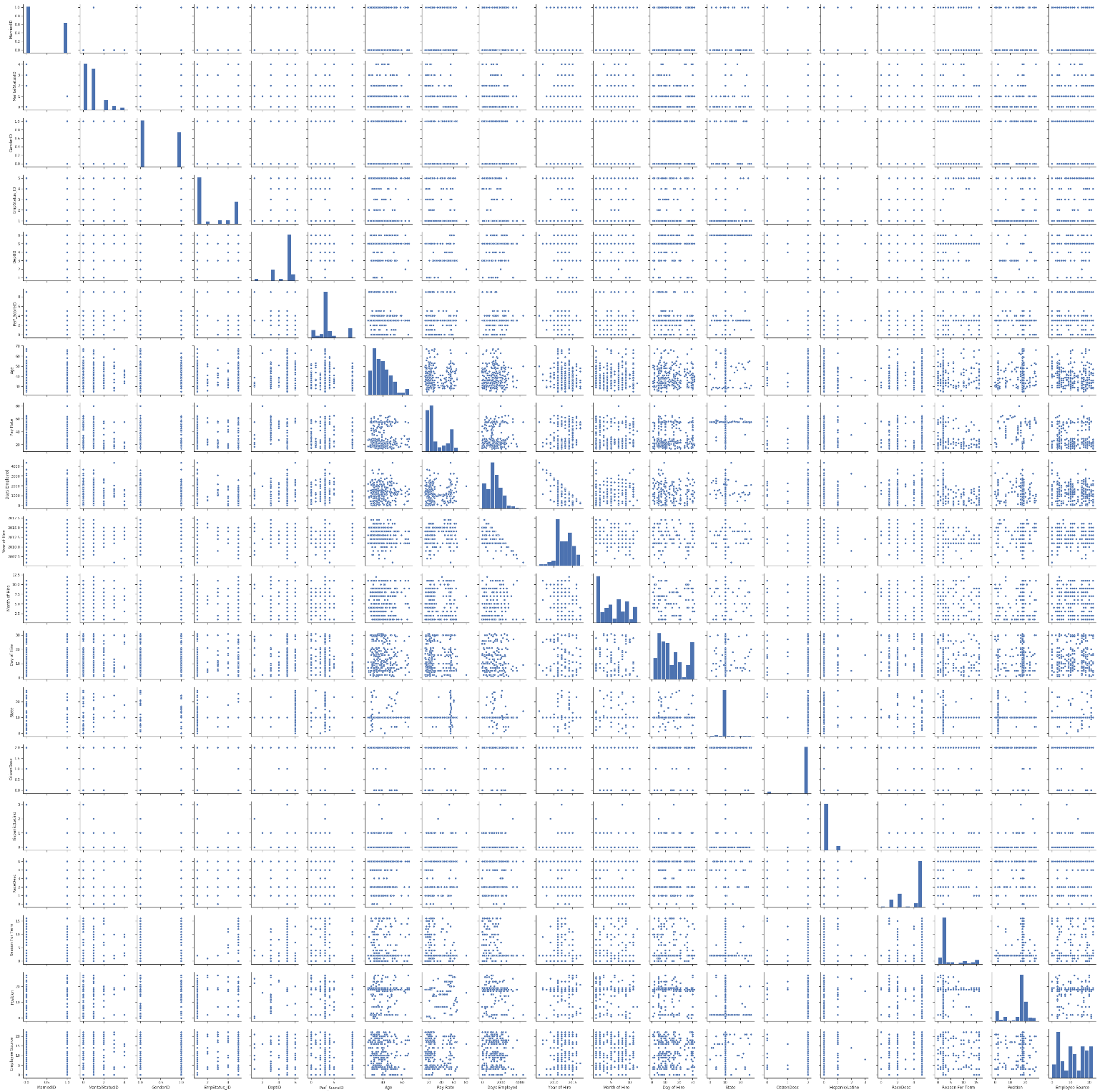


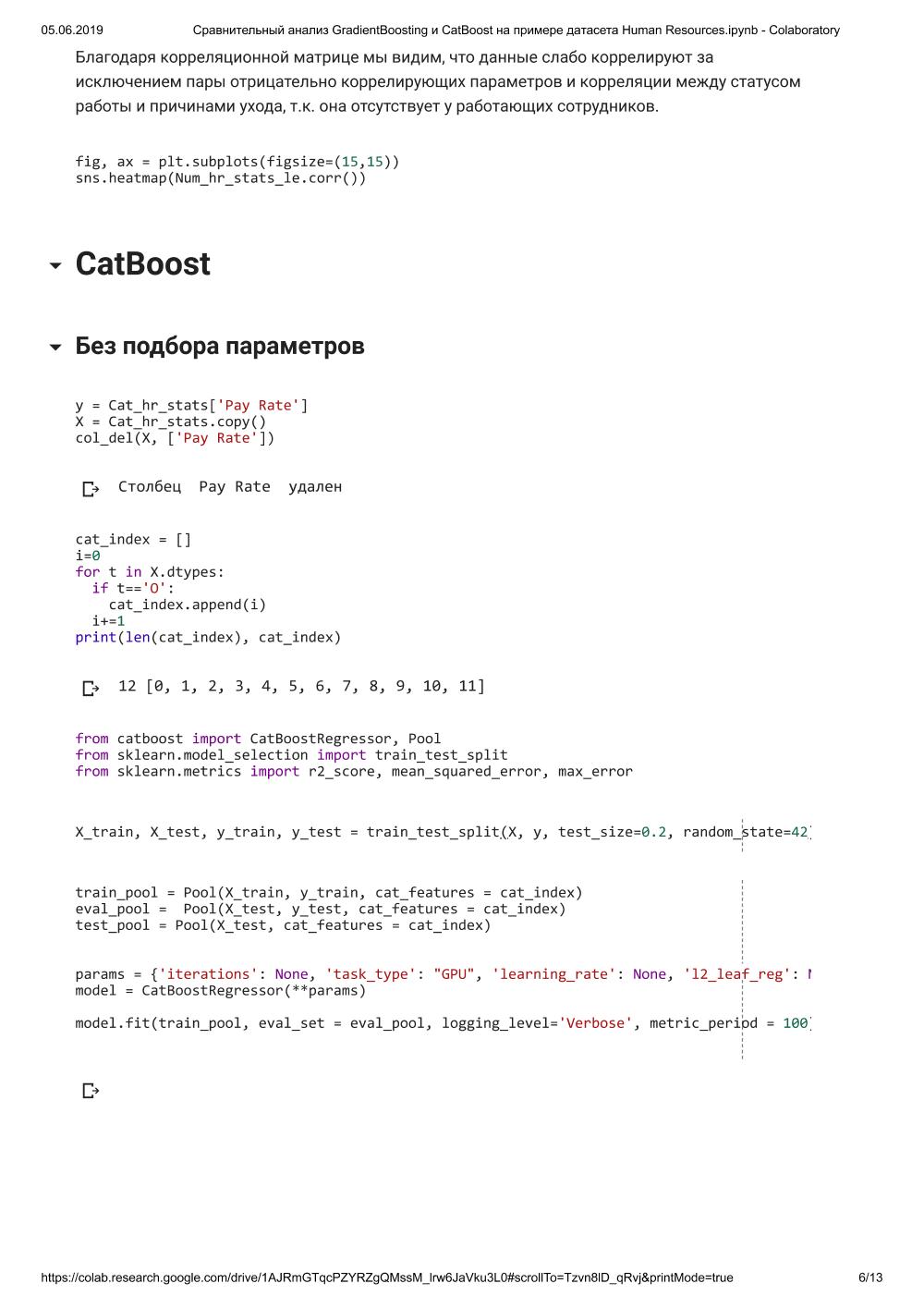


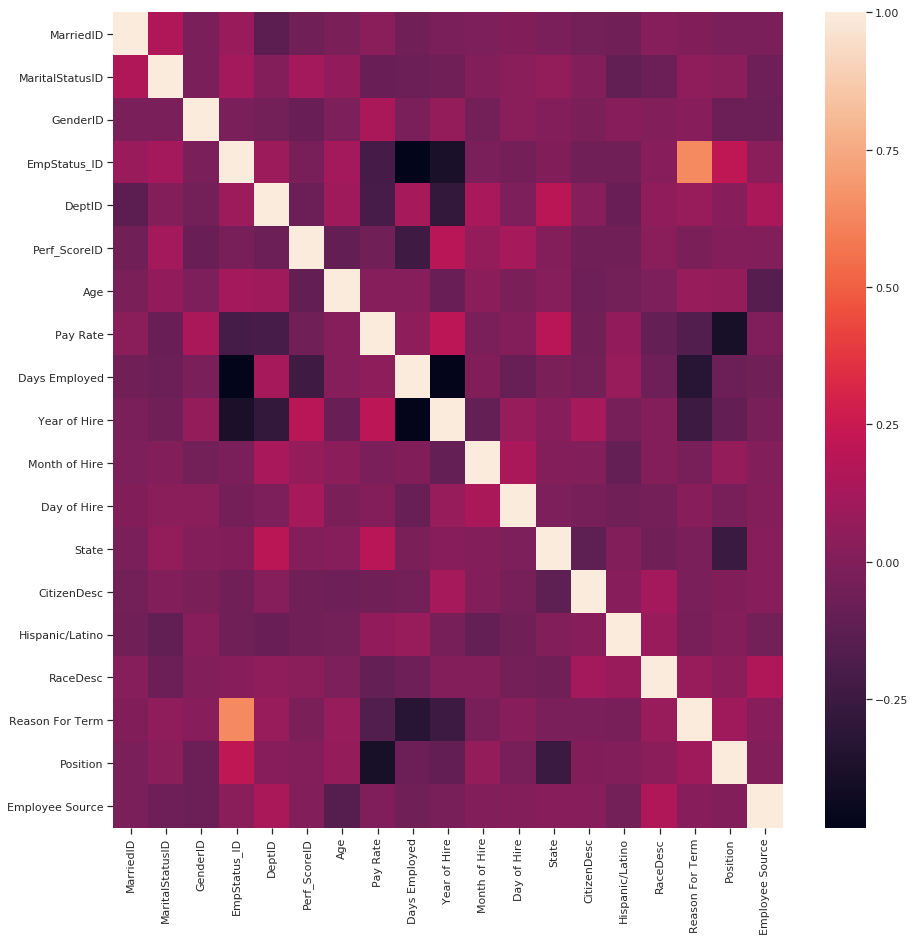




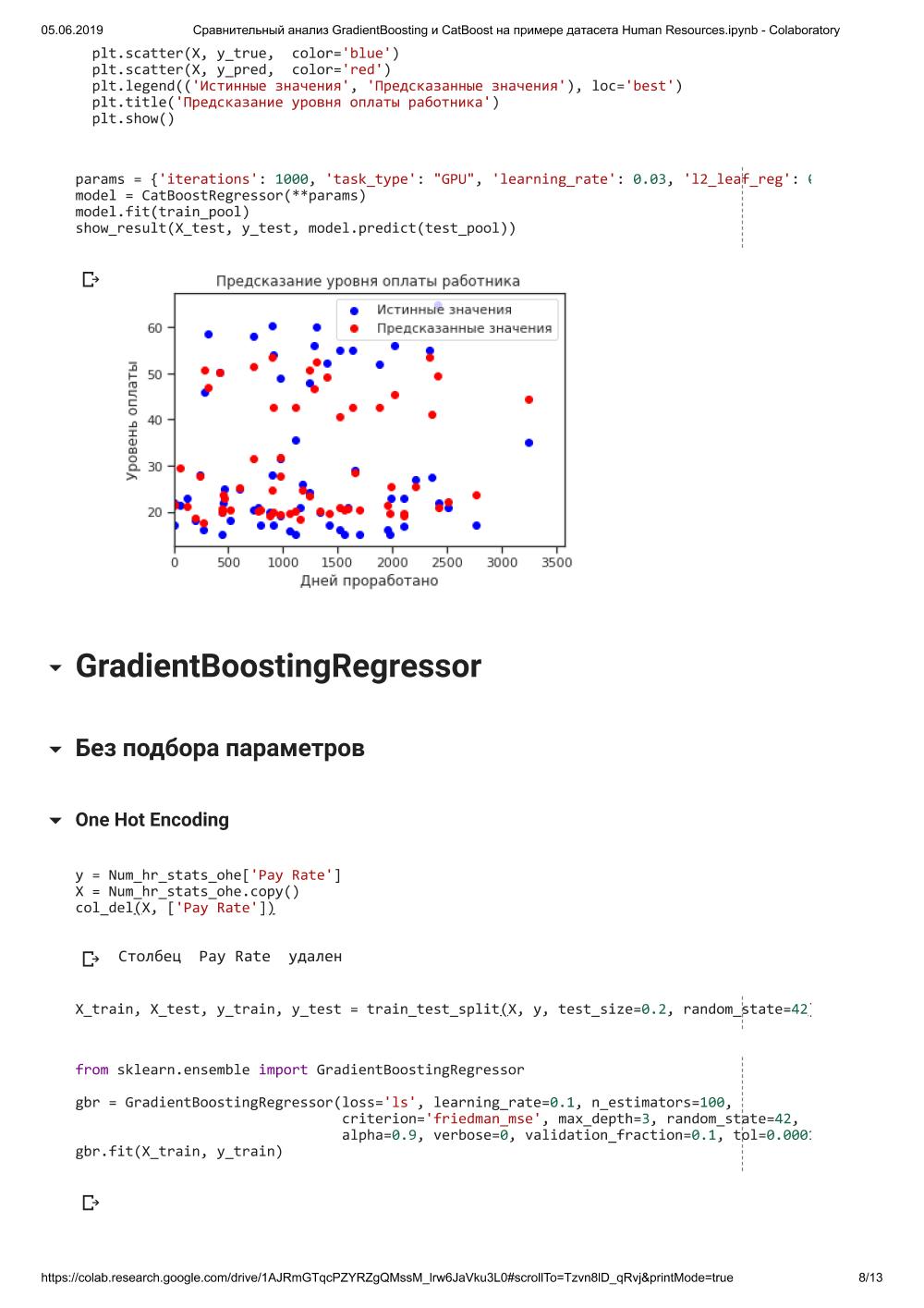


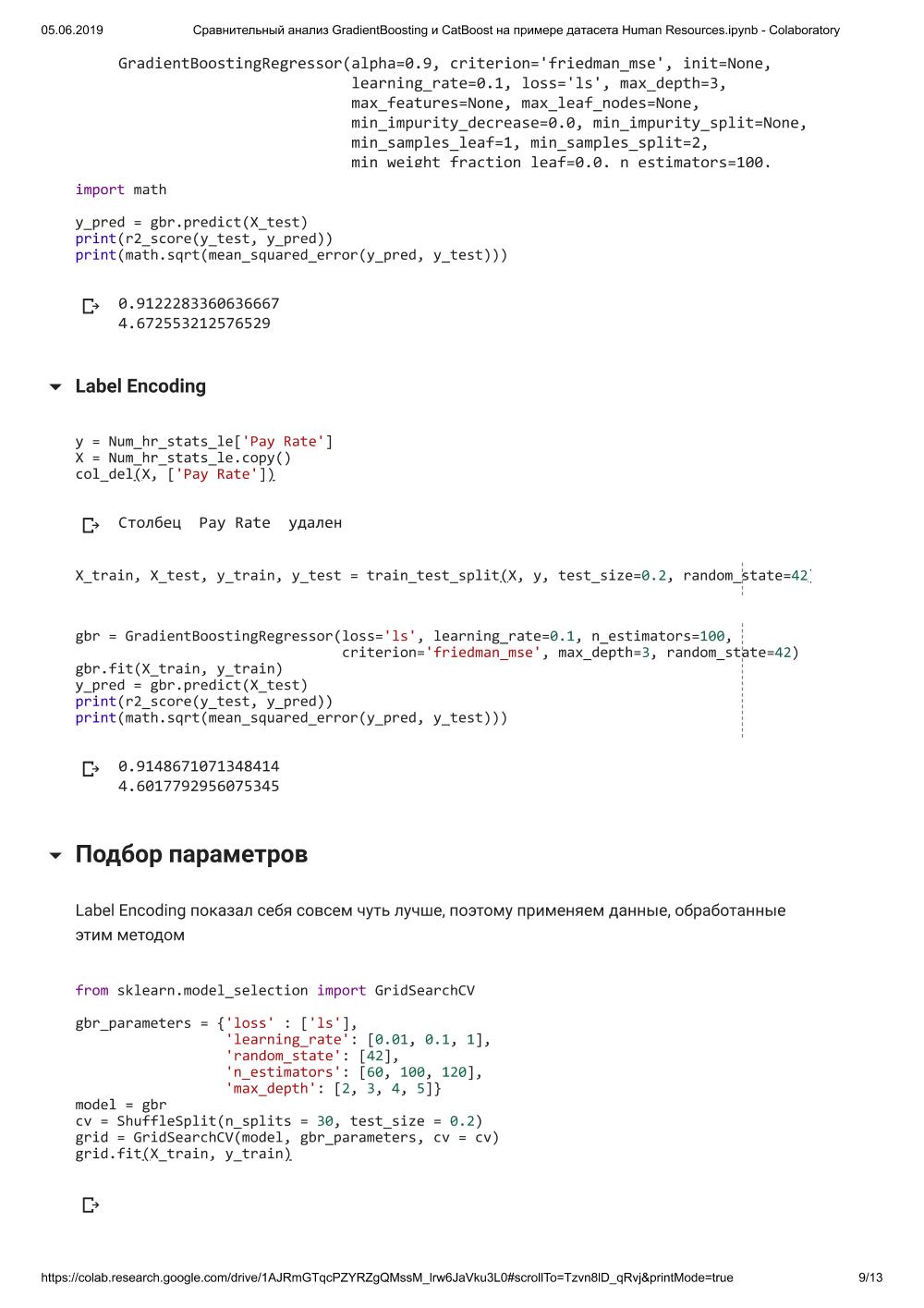




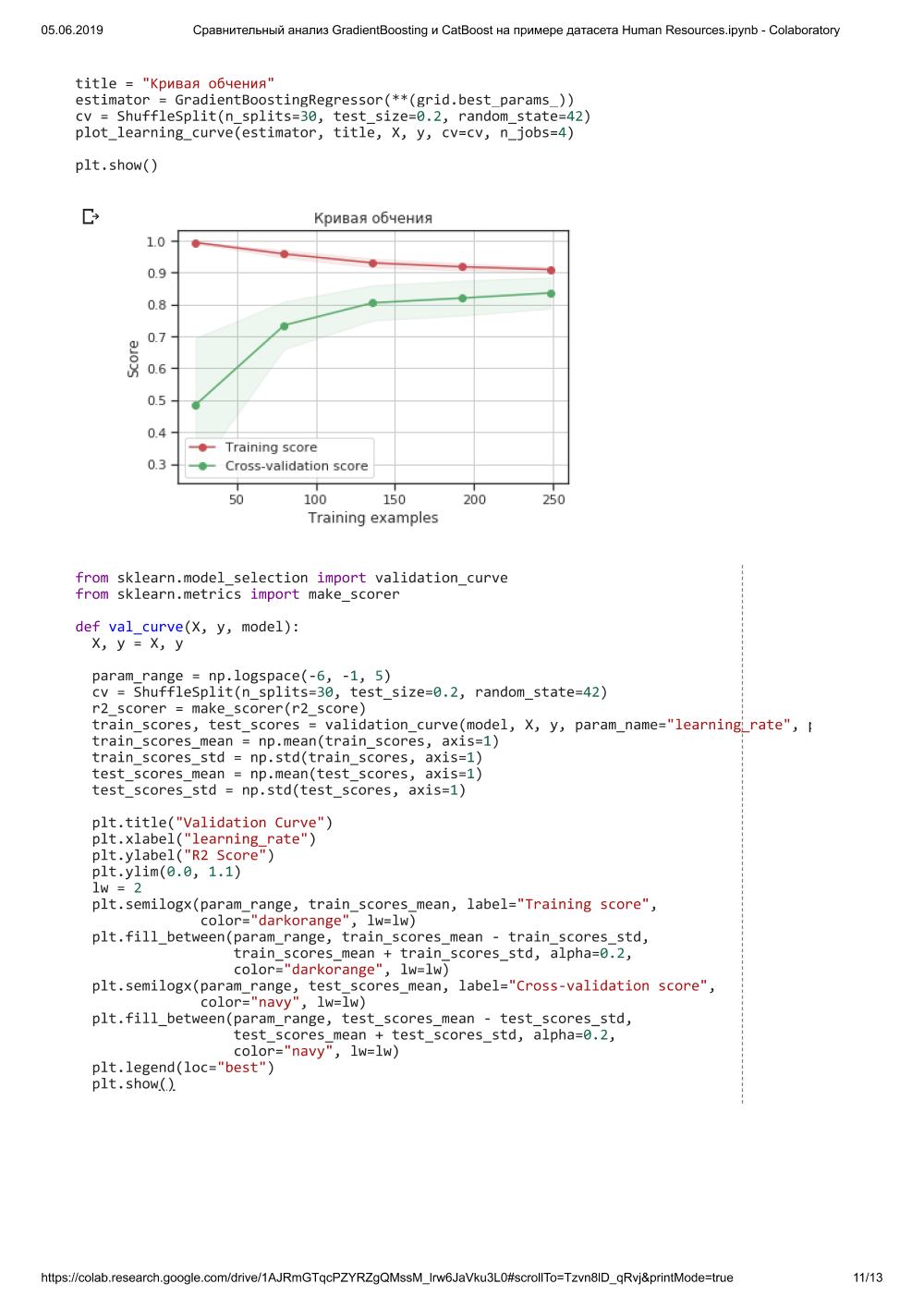


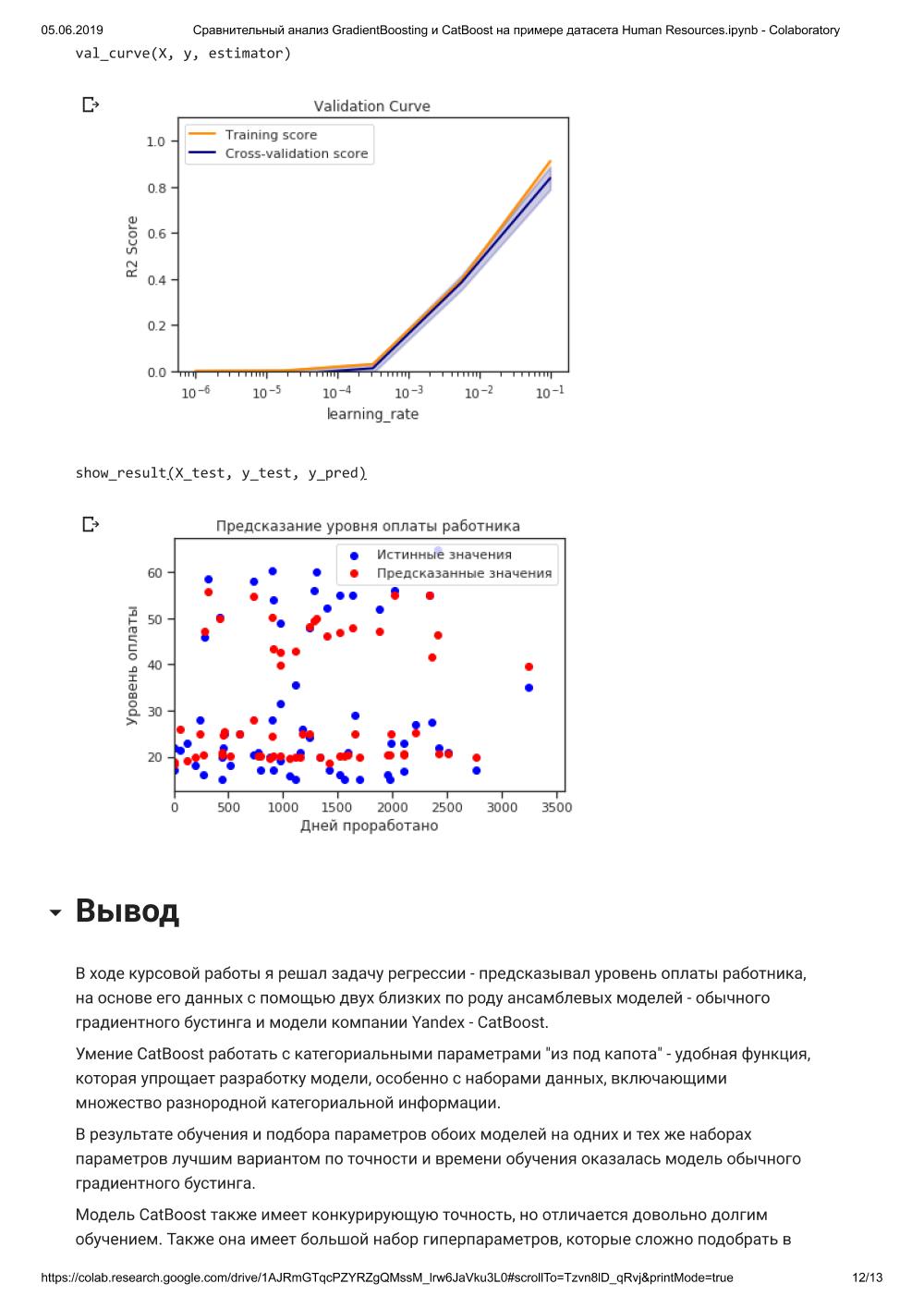












Вывод:

В ходе курсовой работы я решал задачу регрессии - предсказывал уровень оплаты работника, на основе его данных с помощью двух близких по роду ансамблевых моделей - обычного градиентного бустинга и модели компании Yandex - CatBoost.

Умение CatBoost работать с категориальными параметрами "из под капота" - удобная функция, которая упрощает разработку модели, особенно с наборами данных, включающими множество разнородной категориальной информации.

В результате обучения и подбора параметров обоих моделей на одних и тех же наборах параметров лучшим вариантом по точности и времени обучения оказалась модель обычного градиентного бустинга.

Модель CatBoost также имеет конкурирующую точность, но отличается довольно долгим обучением. Также она имеет большой набор гиперпараметров, которые сложно подобрать в рамках курсовой работы. Таким образом, при должной настройке, данный алгоритм наверняка может показать себя с лучшей стороны.

Список литературы:

1. https://scikit-learn.org
2. https://catboost.ai
3. https://ru.wikipedia.org
4. https://github.com/ugapanyuk/ml\_course