|  |  |
| --- | --- |
|  | **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 РТ

КАФЕДРА ИУ5

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К КУРСОВОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***Прогнозирование трансферной***

***стоимости футболистов***

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент РТ5-61 **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Орлова С. М\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсовой работы **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Гапанюк Ю. Е.\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2020 г.*

**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)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

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

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение курсовой работы**

по дисциплине Технологии машинного обучения

Студент группы РТ5-61

Орлова Светлана Михайловна

(Фамилия, имя, отчество)

Тема курсовой работы

Прогнозирование трансферной стоимости футболистов

Направленность КР (учебная, исследовательская, практическая, производственная, др.)

Учебная

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения работы: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

***Задание*** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

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

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

***Оформление курсовой работы:***

Расчетно-пояснительная записка на 16 листах формата А4.

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

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

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

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

Дата выдачи задания « 10 » апреля 2020 г.

**Руководитель курсовой работы**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гапанюк Ю. Е.

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Орлова С. М.

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Содержание

[Введение 4](#_Toc43223653)

[Основная часть 5](#_Toc43223654)

[Импорт библиотек 5](#_Toc43223655)

[Загрузка и первичный анализ данных 5](#_Toc43223656)

[Заполнение пропусков в данных 6](#_Toc43223657)

[Кодирование категориальных признаков 7](#_Toc43223658)

[Масштабирование данных 7](#_Toc43223659)

[Корреляционный анализ данных 8](#_Toc43223660)

[Построение графиков 9](#_Toc43223661)

[Выбор метрик для оценки качества моделей 9](#_Toc43223662)

[Формирование обучающей и тестовой выборок 9](#_Toc43223663)

[Метод ближайших соседей 10](#_Toc43223664)

[Линейная модель 10](#_Toc43223665)

[Метод опорных векторов 12](#_Toc43223666)

[Метод случайного леса 13](#_Toc43223667)

[AdaBoost 13](#_Toc43223668)

[Сравнение качества построенных моделей 14](#_Toc43223669)

[Заключение 15](#_Toc43223670)

[Список использованных источников 16](#_Toc43223671)

Введение

Курсовой проект – самостоятельная часть учебной дисциплины «Технологии машинного обучения» – учебная и практическая исследовательская студенческая работа, направленная на решение комплексной задачи машинного обучения. Результатом курсового проекта является отчет, содержащий описания моделей, тексты программ и результаты экспериментов.

Курсовой проект опирается на знания, умения и владения, полученные студентом в рамках лекций и лабораторных работ по дисциплине.

В рамках курсового проекта проводится типовое исследование. Типовое исследование – решение задачи машинного обучения на основе материалов дисциплины. Выполняется студентом единолично.

Схема типового исследования, проводимого студентом в рамках курсовой работы, содержит выполнение следующих шагов:

* Поиск и выбор набора данных для построения моделей машинного обучения. На основе выбранного набора данных необходимо построить модели машинного обучения для решения задачи регрессии.
* Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.
* Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
* Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
* Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее трех метрик.
* Выбор наиболее подходящих моделей для решения задачи регрессии. Необходимо использовать не менее пяти моделей, две из которых должны быть ансамблевыми.
* Формирование обучающей и тестовой выборок на основе исходного набора данных.
* Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки.
* Подбор гиперпараметров для выбранных моделей. Рекомендуется использовать методы кросс-валидации.
* Повторение построение базового решения для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.
* Формирование выводов о качестве построенных моделей на основе выбранных метрик. Результаты сравнения качества рекомендуется отобразить в виде графиков и сделать выводы в форме текстового описания.

Приведенная схема исследования является рекомендуемой. В зависимости от решаемой задачи возможны модификации.

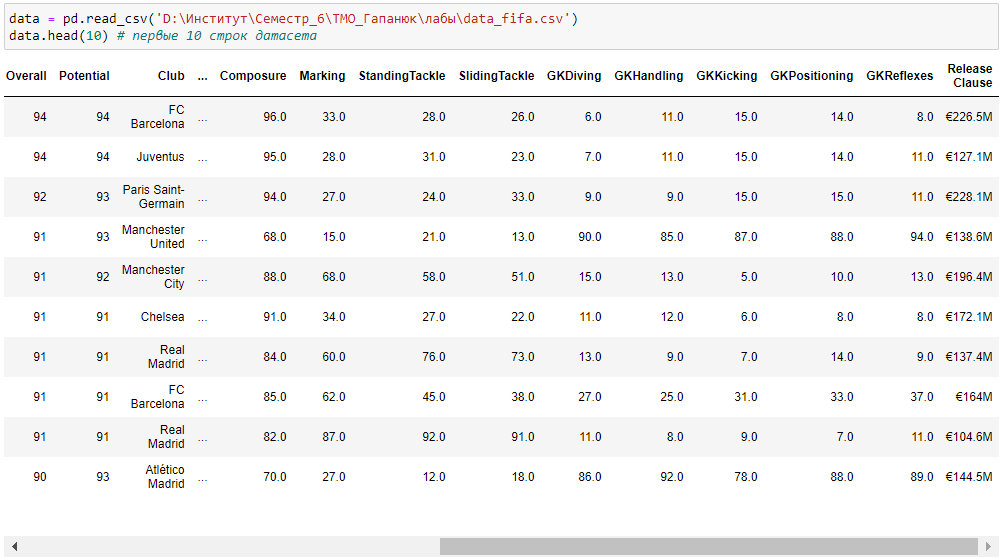
В данной работе используется датасет https://www.kaggle.com/karangadiya/fifa19, в котором описаны физические и технические данные футболистов. Целевая функция Release Clause показывает трансферную стоимость футболистов.

Основная часть

Импорт библиотек



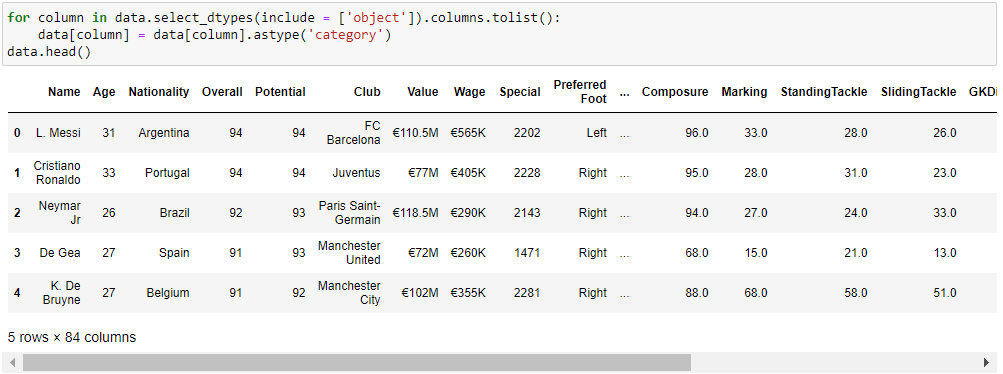
Загрузка и первичный анализ данных

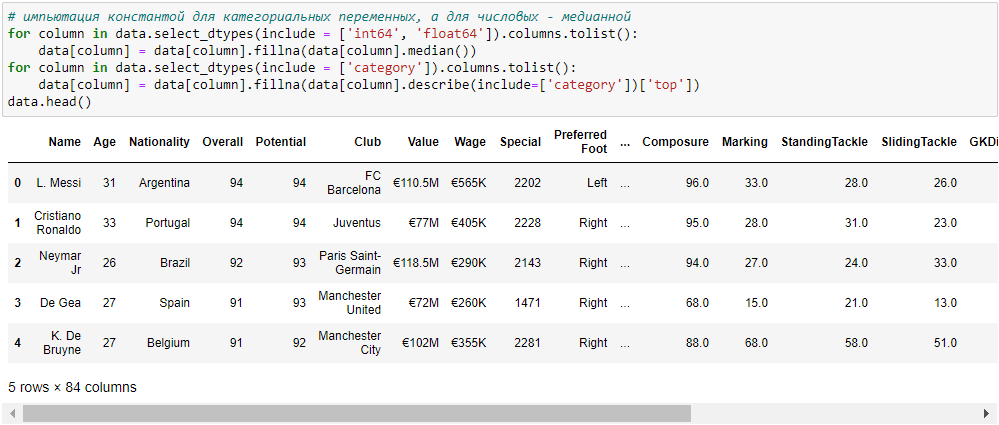




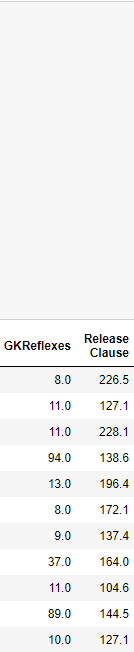
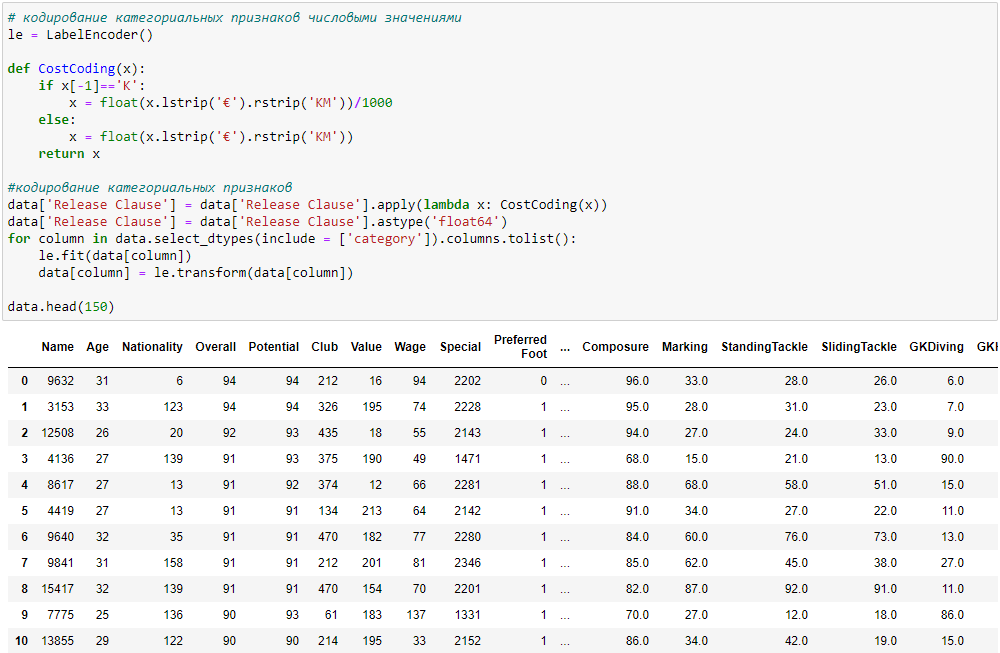


Заполнение пропусков в данных

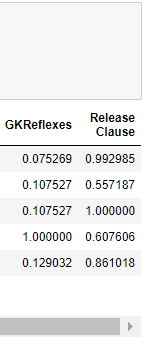
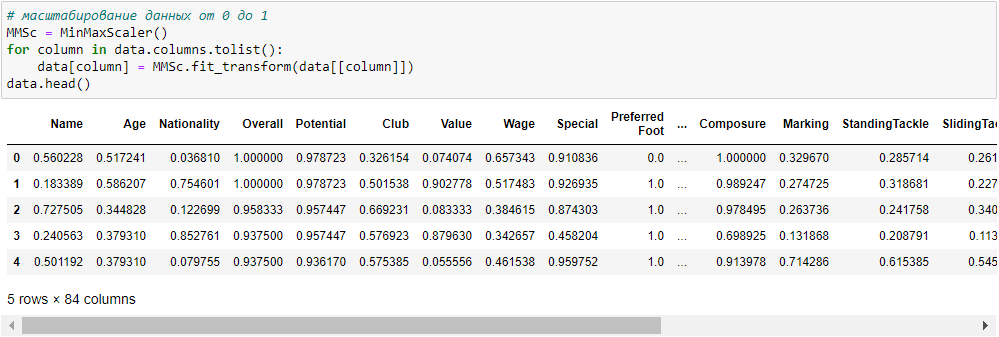




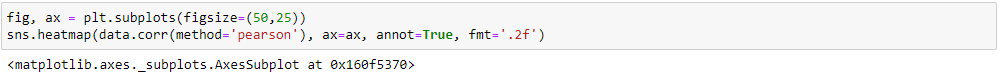
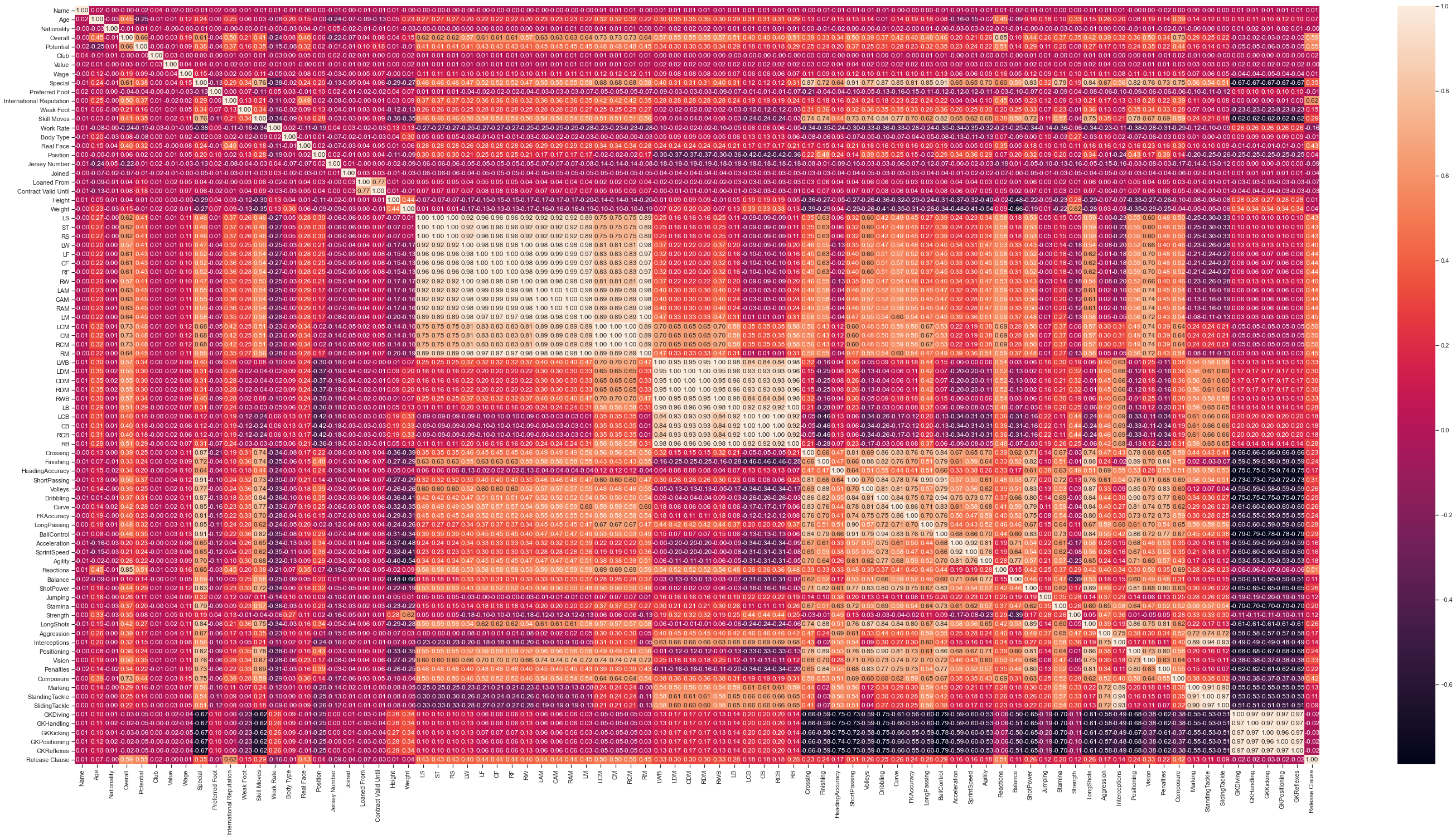
Кодирование категориальных признаков



Масштабирование данных



Корреляционный анализ данных

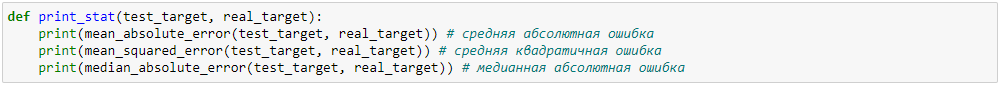
  




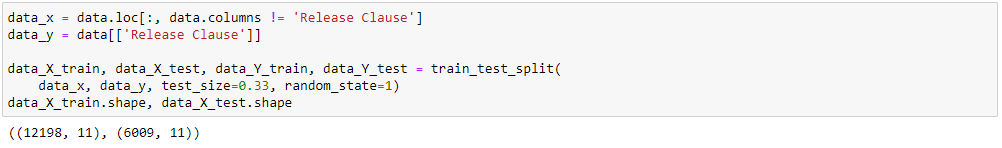
Построение графиков

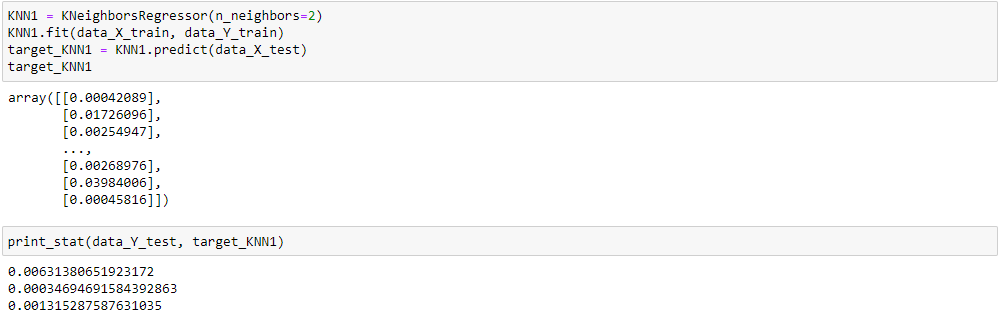

Выбор метрик для оценки качества моделей

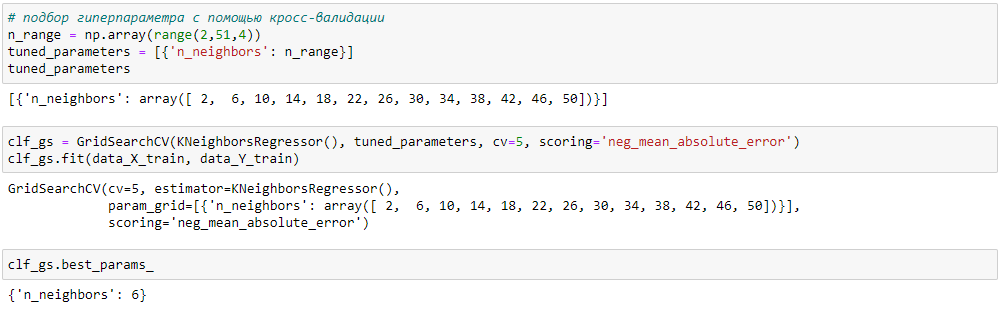


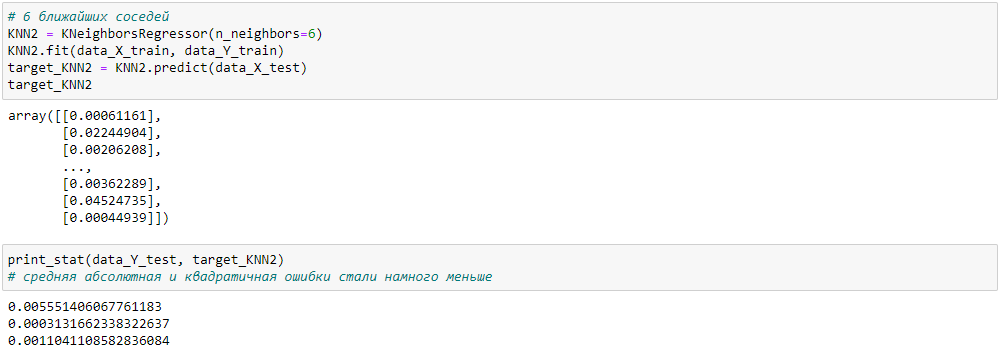
Формирование обучающей и тестовой выборок



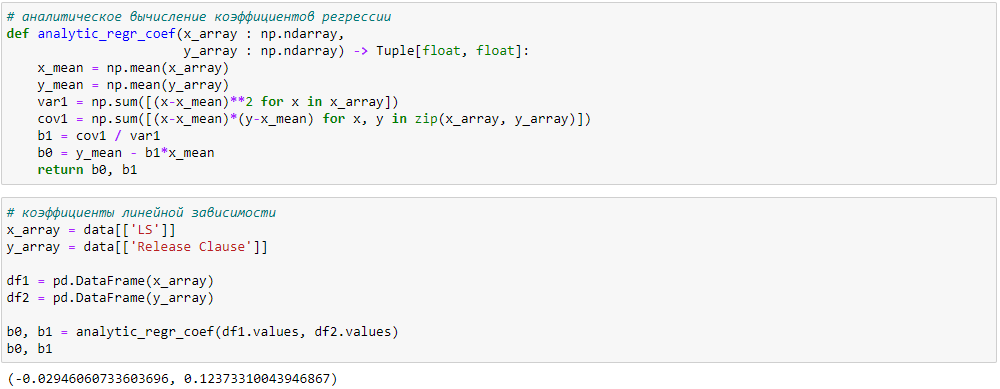
Метод ближайших соседей

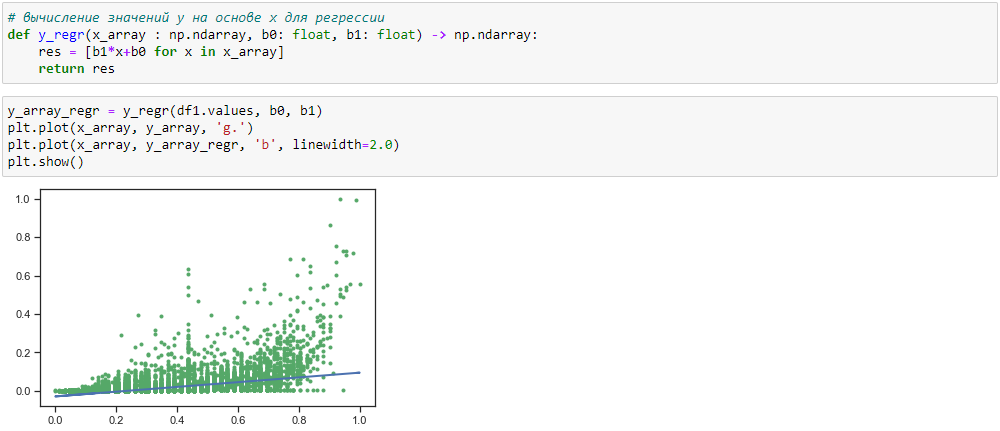


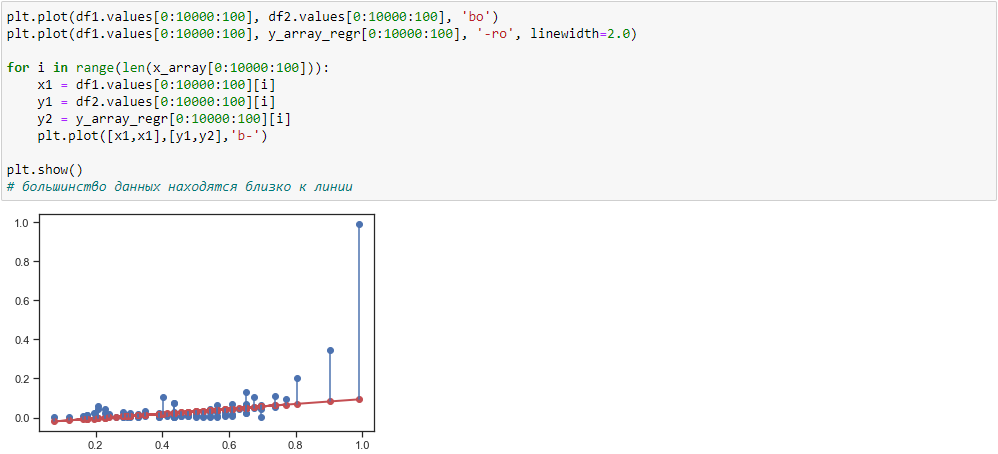


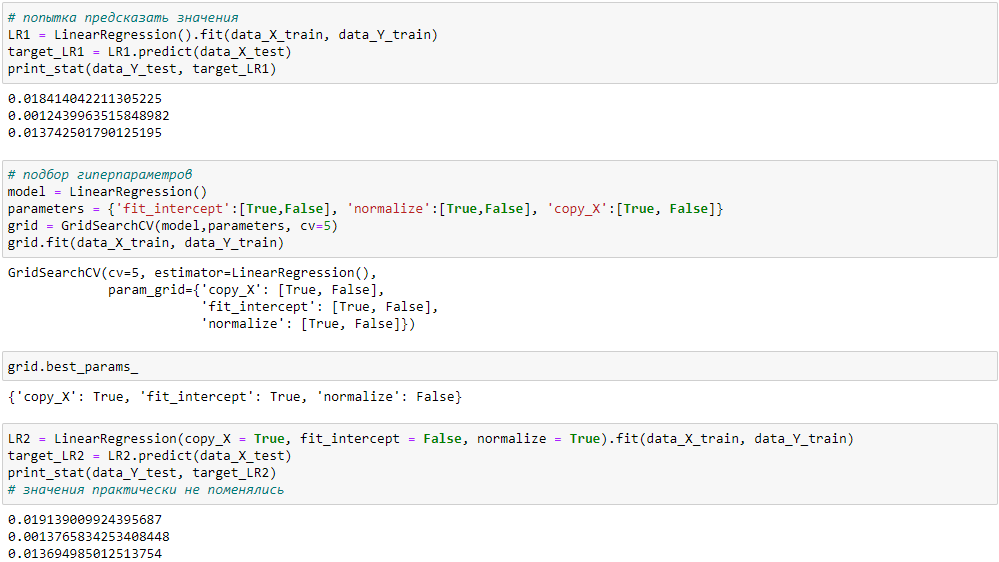


Линейная модель

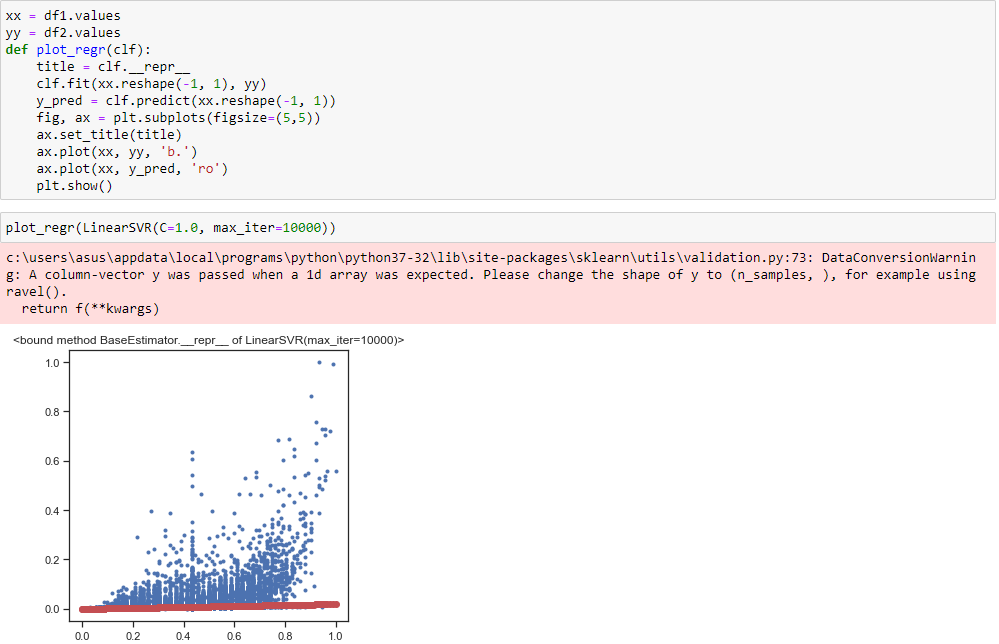


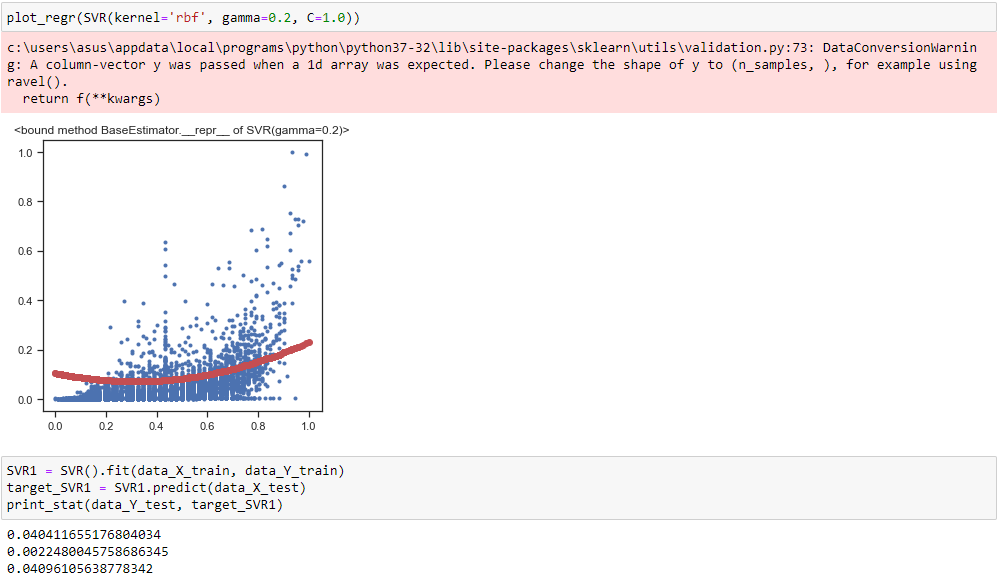


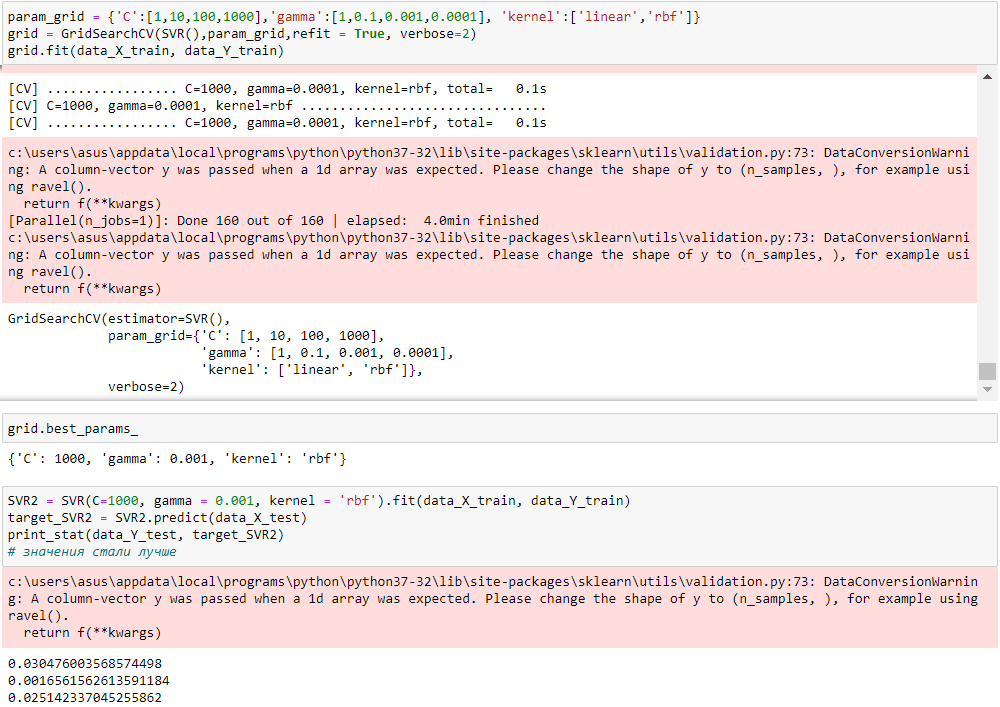


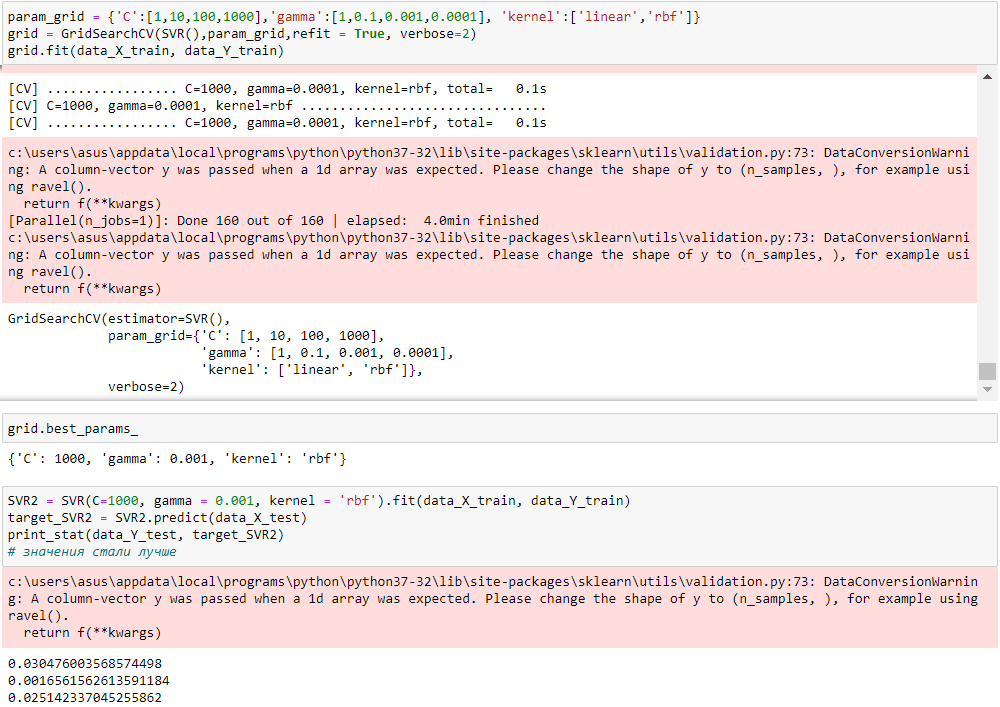


Метод опорных векторов

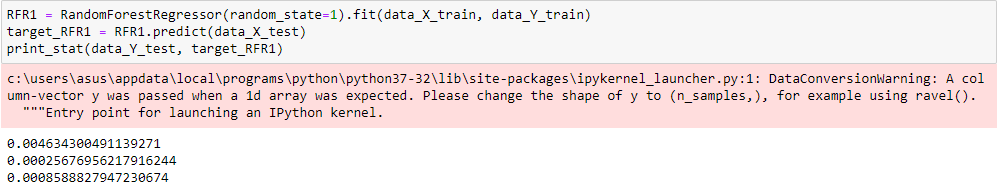


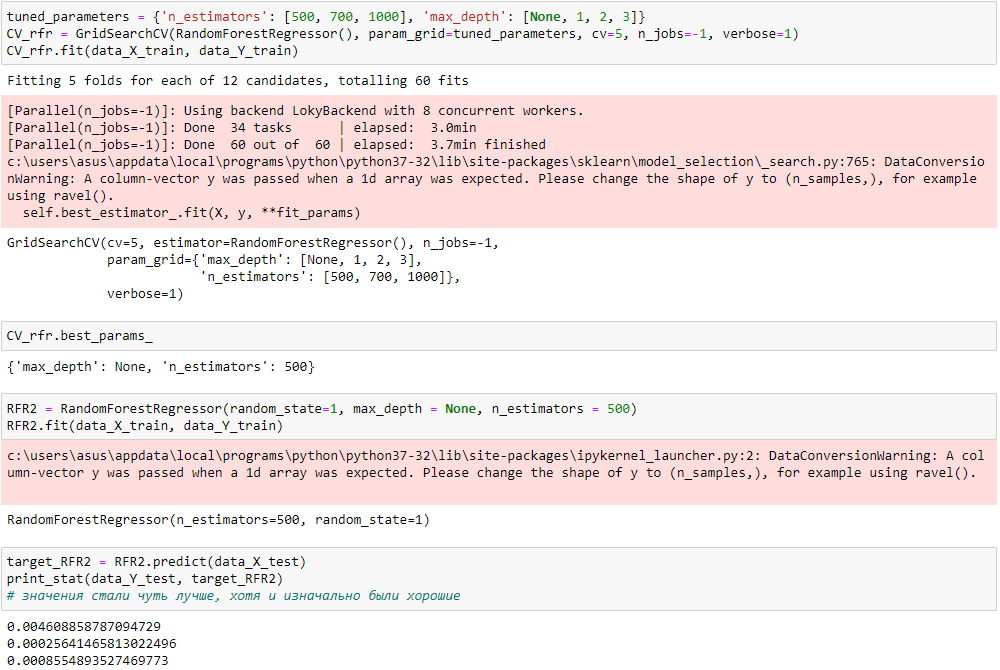




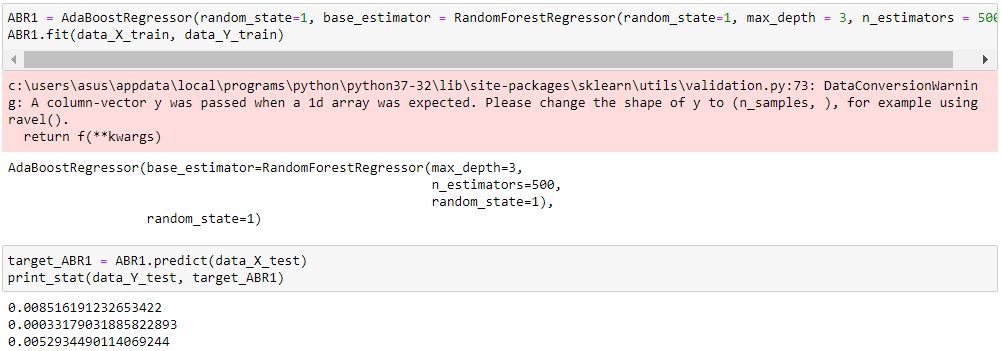


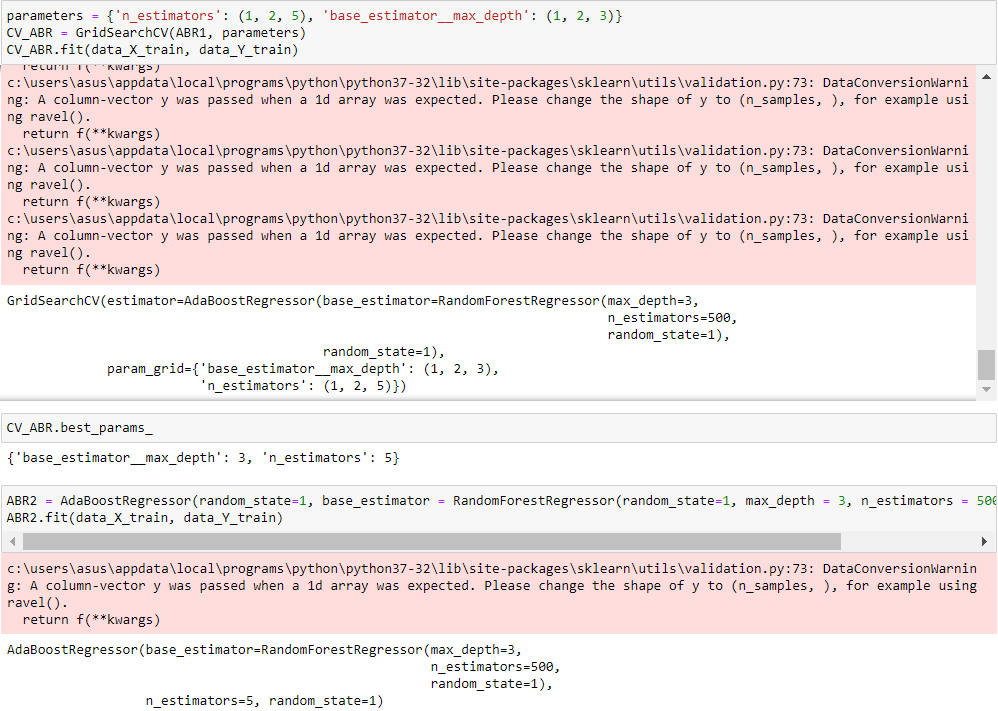
Метод случайного леса

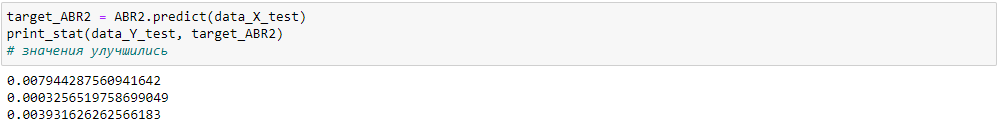




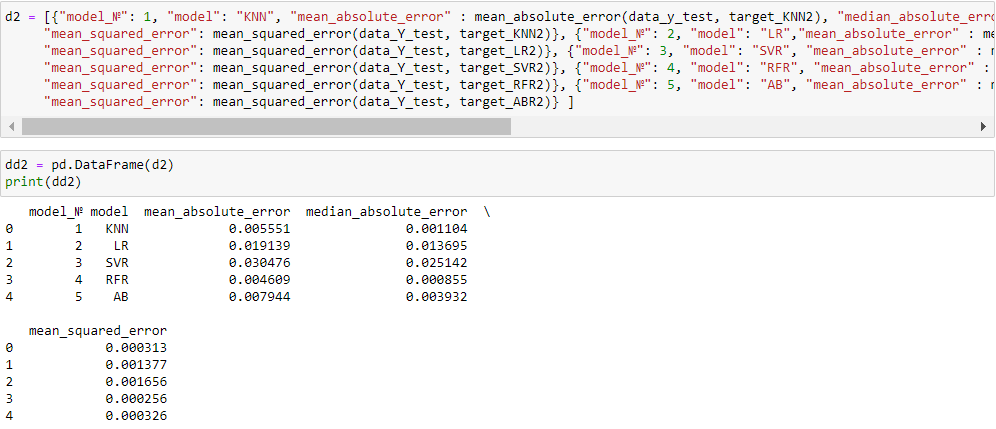
AdaBoost







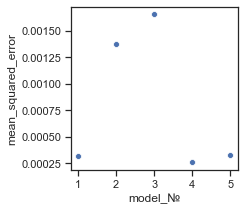
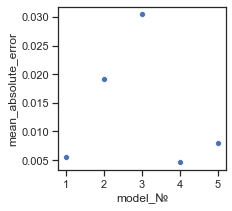
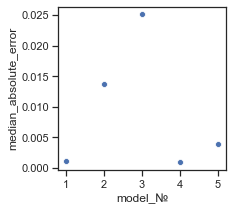
Сравнение качества построенных моделей

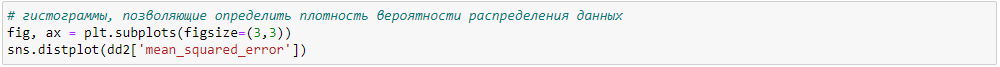






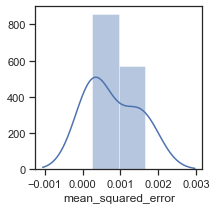
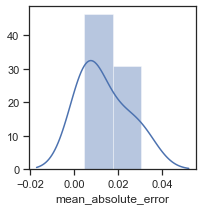
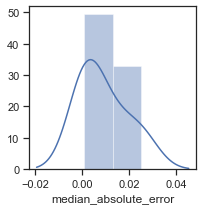






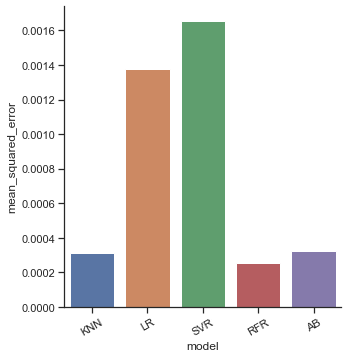
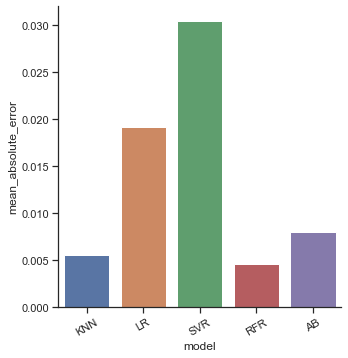
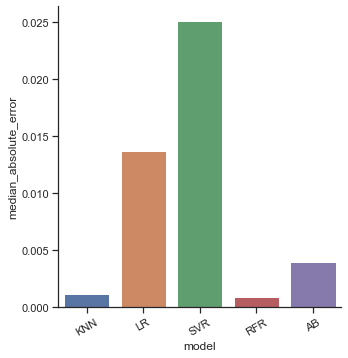








При анализе метрик почти всегда лидирует Random Forest Regressor модель. Для данного датасета лучше всего использовать этот метод. Так же можно заметить, что подбор гиперпараметров почти всегда улучшает метрики.

Заключение

В данной курсовой работе мы научились работать с различными моделями данных и проводить их анализ. Был проведен анализ данных футболистов и были использованы различные модели машинного обучения для предсказания трансферной стоимости. Можно сделать вывод, что, зная технические характеристики и рейтинг футболиста, предсказать его стоимость можно с высокой точностью.

Список использованных источников

* <https://habr.com/ru/post/279117/>
* <https://scikit-learn.org/stable/modules/grid_search.html>
* <http://www.williamspublishing.com/PDF/978-5-9500296-2-2/part.pdf>
* Лекции по курсу ТМО