МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Санкт-Петербургский национальный исследовательский университет

информационных технологий, механики и оптики»

Факультет информационных технологий и программирования

Кафедра информационных систем

Лабораторная работа №4

**Обучение модели**

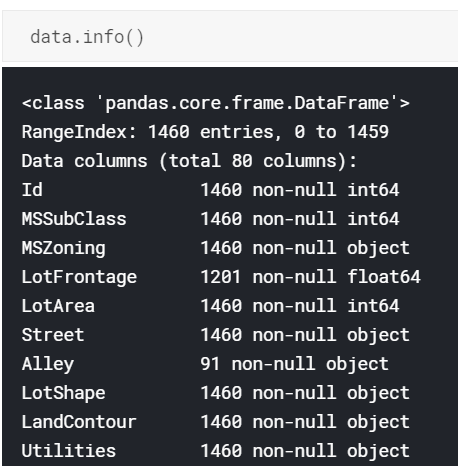
Выполнил студент группы № М3307:   
Бойцов Виталий Вячеславич

Санкт-Петербург

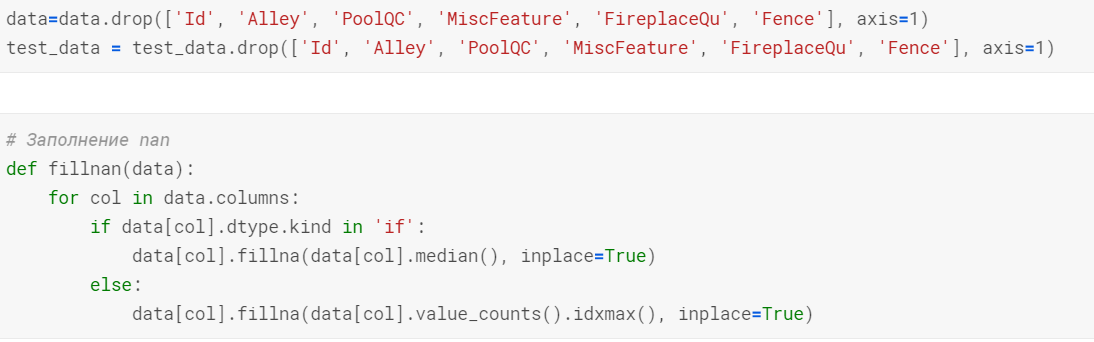
2018

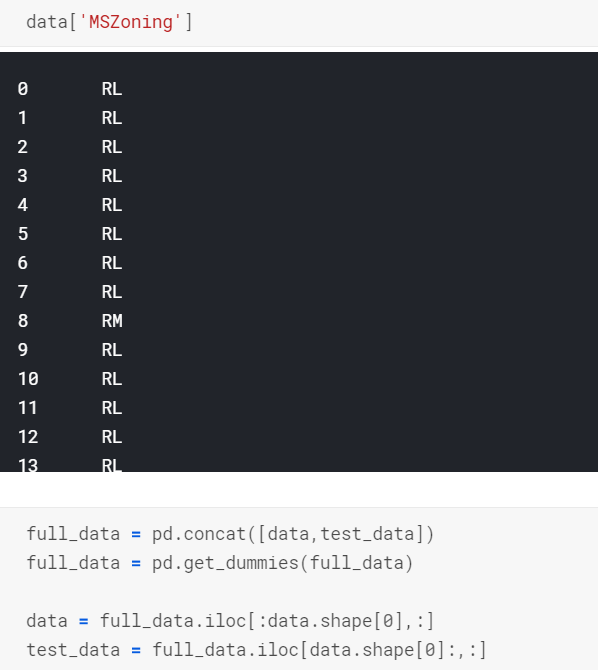
Подготовка данных

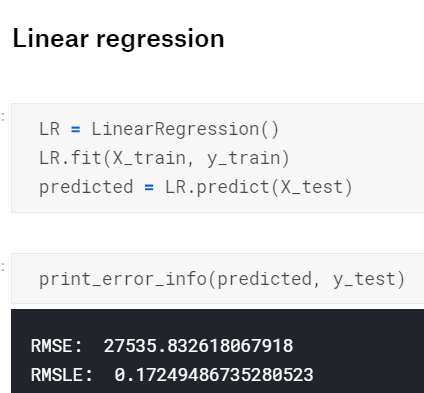
Для начала нам необходимо избавиться от именованных полей, а также обработать поля, допускающие NULL.

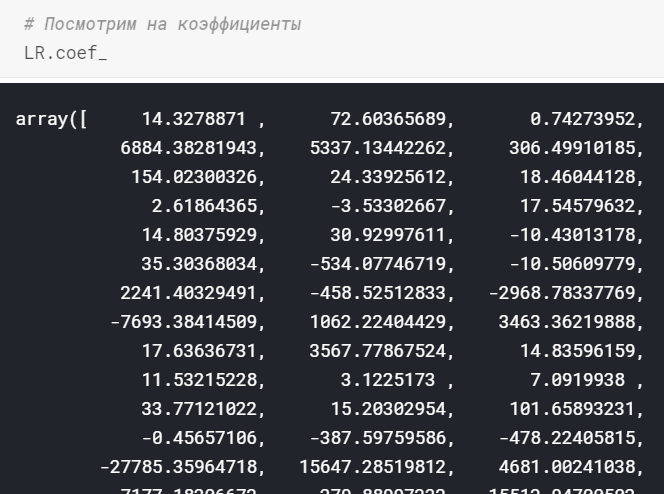


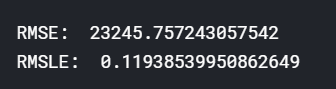
Избавимся от объектов, которые не имеют достаточной информации, т.к. содержат не NULL значение. Остальные Null объекты заполним медианой.

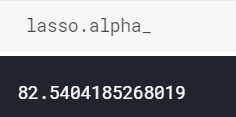


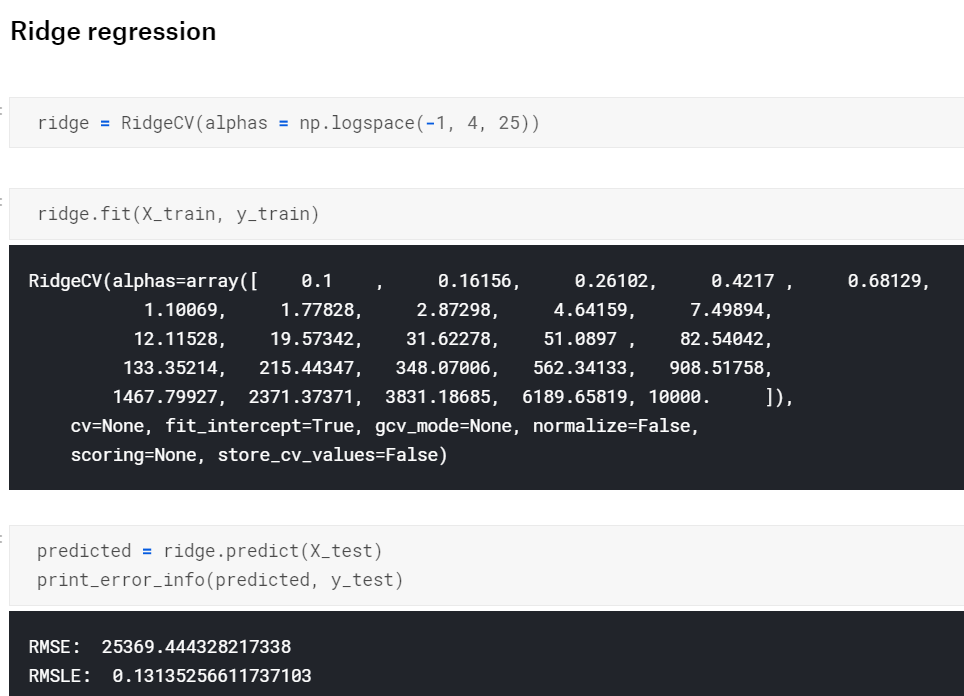
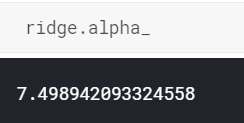
Для именованных полей используем метод get\_dummies()  


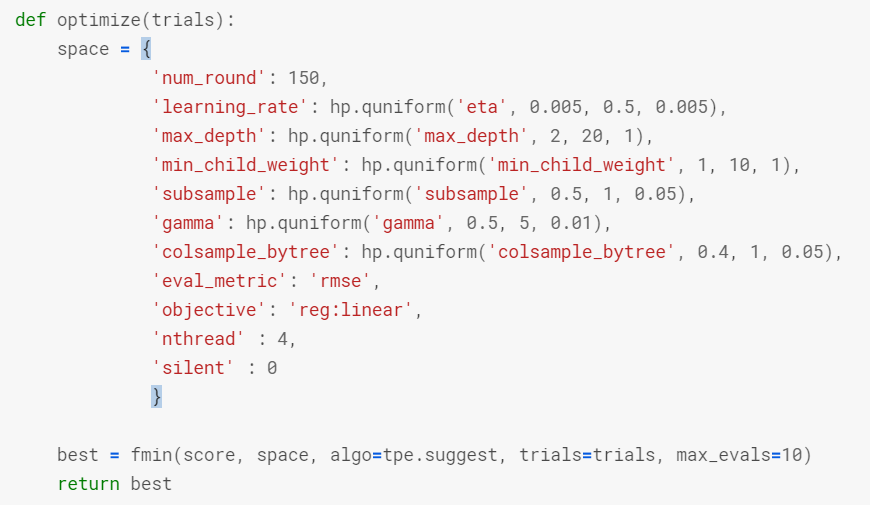
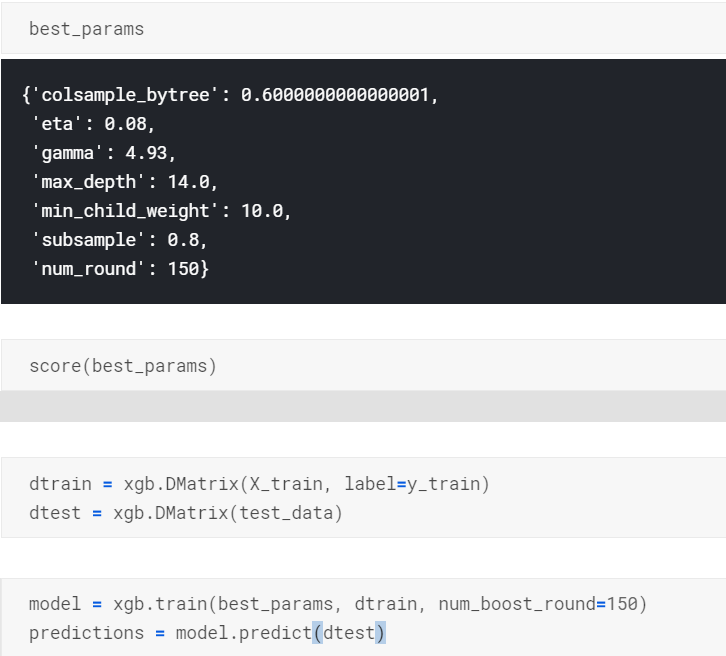






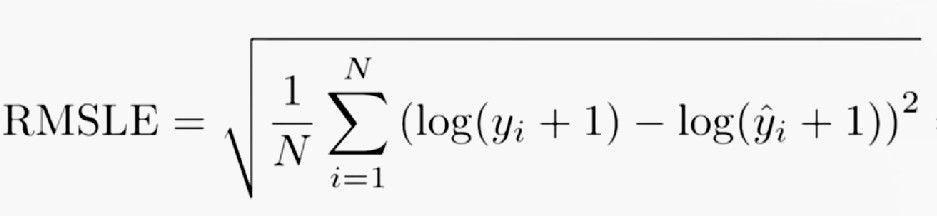
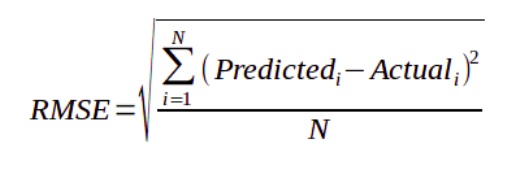
  


XGBoost 

1) Линейная регрессия.  
Способ получения зависимости target feature, заключающийся в получении функции имеющей линейный вид. Каждая фича имеет свой коэффициент, который высчитывается таким образом, чтобы его вклад был более точен.

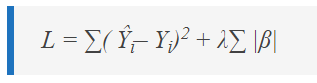
2) Полиномиальная регрессия.  
Способ получения зависимости таргет фичи, заключающийся в получении функции имеющий полиномиальный вид. Многочлен состоит из одночленов n-го порядка, где n – кол-во фичей.

3) Критерий качества, функция ошибки в линейных моделях.  
Функции ошибки позволяют оценивать точность полученных результатов, для этого используются Root Mean Square Error и Root Mean Square Logarithmic Error. Первая метрика используется для оценки абсолютной погрешности, а вторая для относительной.



4) Категориальные признаки  
Категориальные признаки – это фичи, которые имеют именованную шкалу измерения. Для того, чтобы учитывать их используют несколько способов. Можно ставить в соответствие каждому виду какое-то число и заменить фичу с названиями на уникальные числа, однако лучше использовать метод Dummy. Идея этого метода заключается в том, чтобы разбить нужную нам фичу на несколько и ставить 1 в том столбце, которому принадлежит данный объект.

5) Оптимизация гиперпараметров  
Линейную регрессию можно улучшить, введя дополнительные параметры. Например, с помощью параметра нулевого порядка можно задать функцию, которая не будет проходить через начало координат. Lasso регрессия добавляет гиперпараметр – alpha.



6) Метод наименьших квадратов – способ решения различных задач, основанный на минимизации суммы квадратов отклонений некоторой функции