|  |  |
| --- | --- |
| 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»**

Курсовая работа

по дисциплине «Технологии машинного обучения»

Пояснительная записка

(вид документа)

писчая бумага

(вид носителя)

12

(количество листов)

Выполнил:

студентка группы ИУ5-64

Изъюрова Маргарита

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

подпись, дата

Проверил:

Гапанюк Юрий Евгеньевич

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

подпись, дата

2019 г.

Содержание:

1. [Задание 3](#_Toc6264189)
2. [Введение 3](#_Toc6264194)
3. Постановка задачи 4
4. Описание выбранного датасета 4
5. Текст программы с примерами выполнения (листинг) 5
6. Заключение 12
7. Список литературы 12

## Задание

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

1. Поиск и выбор набора данных для построения моделей машинного обучения. На основе выбранного набора данных студент должен построить модели машинного обучения для решения или задачи классификации, или задачи регрессии.
2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.
3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. В зависимости от набора данных, порядок выполнения пунктов 2, 3, 4 может быть изменен.
5. Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее трех метрик и обосновать выбор.
6. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. Необходимо использовать не менее пяти моделей, две из которых должны быть ансамблевыми.
7. Формирование обучающей и тестовой выборок на основе исходного набора данных.
8. 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.
9. Подбор гиперпараметров для выбранных моделей. Рекомендуется использовать методы кросс-валидации. В зависимости от используемой библиотеки можно применять функцию GridSearchCV, использовать перебор параметров в цикле, или использовать другие методы.
10. Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.
11. 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. Рекомендуется постройение графиков обучения и валидации, влияния значений гиперпарметров на качество моделей и т.д.

Приведенная схема исследования является рекомендуемой. Возможно выполнение курсовой работы на нестандартную тему, которая должна быть предварительно согласована с ответственным за прием курсовой работы.

**Введение**

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

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

В рамках данной курсовой работы необходимо применить навыки, полученные в течение курса «Технологии машинного обучения», и обосновать полученные результаты.

**Постановка задачи**

Сейчас у каждого более-менее известного бизнеса есть свое приложение, либо сайт. С помощью данных сайтов и приложений можно грамотно выстраивать рекламу внутри своего продукта. Для того, чтобы реклама предлагалась действительно грамотно, необходимо понимать, какие у нас есть пользователи, и что может заинтересовать каждого из них.

Для этого необходимо собирать определенную информацию о каждом пользователе. Так как предполагается, что в приложении имеется доступ к пользовательским действиям, мы можем знать, какими функциями нашего продукта пользуется наш клиент, и предлагать эти функции другим клиентам направленно.

Я взяла датасет, с помощью которого в теории можно решить задачу.

**Описание выбранного датасета**

Для выполнения курсовой работы я выбрала датасет "Mobile App Statistics (Apple iOS app store)"

<https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps/downloads/app-store-apple-data-set-10k-apps.zip/7>

Постоянно меняющийся мобильный ландшафт - это сложное пространство для навигации. Процент мобильных относительно стационарных только растет. Android занимает около 53,2% рынка смартфонов, а iOS - 43%. Чтобы заставить больше людей загружать ваше приложение, вам нужно убедиться, что они могут легко найти ваше приложение. Аналитика мобильных приложений - отличный способ понять существующую стратегию, способствующую росту и удержанию будущего пользователя. В настоящее время с миллионами приложений следующий набор данных стал очень важным для получения самых популярных приложений в магазине приложений iOS. Этот набор данных содержит более 7000 деталей мобильного приложения Apple iOS. Данные были извлечены из API поиска iTunes на веб-сайте Apple Inc. Для этого исследования использовались R и Linux инструменты для утилизации.

Датасет содержит следующие поля:

"id" : ID приложения (не используется в модели)

"track\_name": название приложения (не используется в модели)

"size\_bytes": размер в байтах

"currency": тип валюты (не используется в модели)

"price": цена

"rating\_count\_tot": счетчик оценок пользователей (для всех версий)

"rating\_count\_ver": счетчик оценок пользователей (для текущей версии) (не используется в модели)

"user\_rating" : среднее значение оценок пользователей (для всех версий)

"user\_rating\_ver": среднее значение оценок пользователей (для текущей версии) (не используется в модели)

"ver" : код последней версии (не используется в модели)

"cont\_rating": рейтинг контента (не используется в модели)

"sup\_devices.num": число поддерживаемых устройств

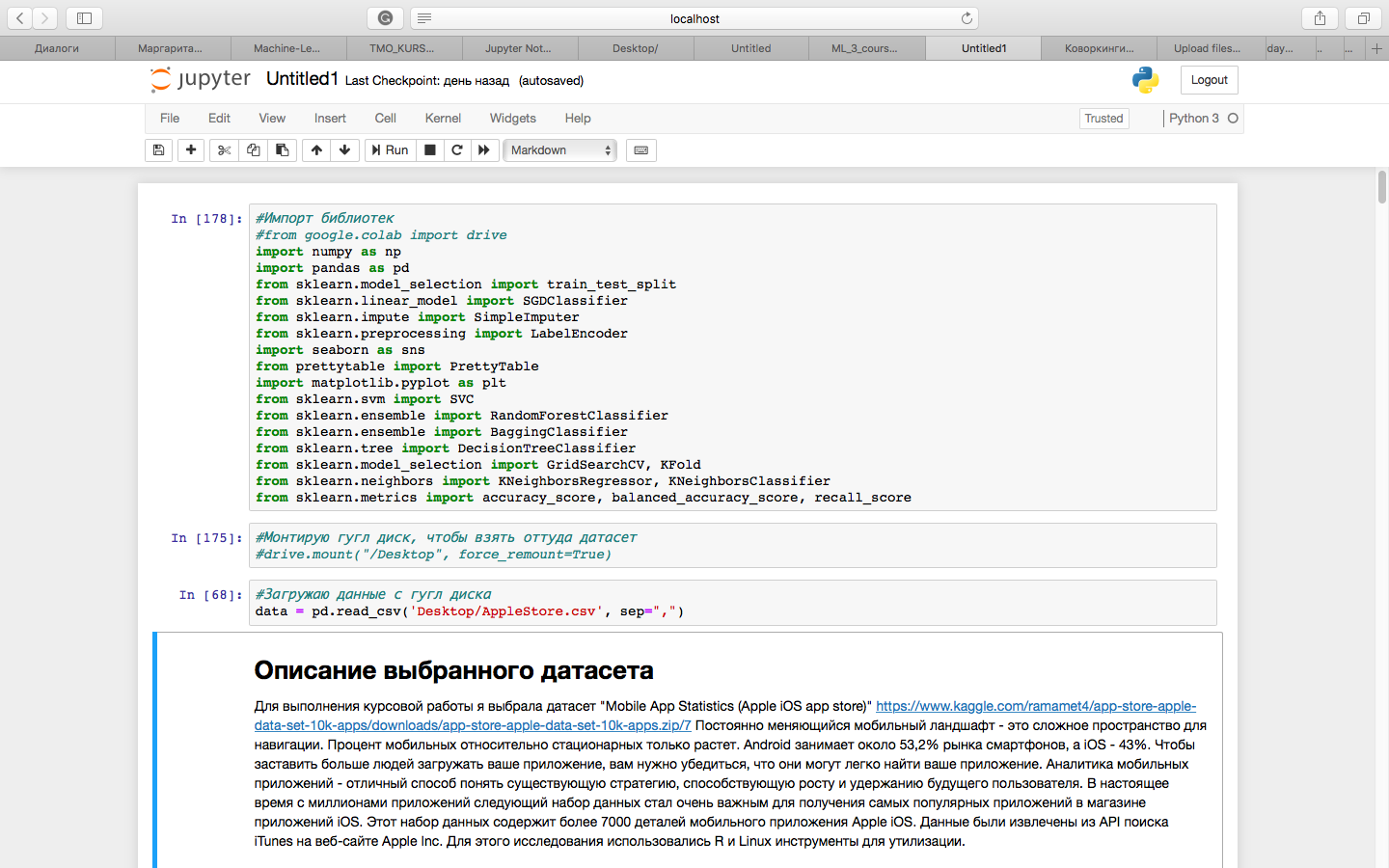
"ipadSc\_urls.num": количество отображаемых снимков экрана

"lang.num": число поддерживаемых языков

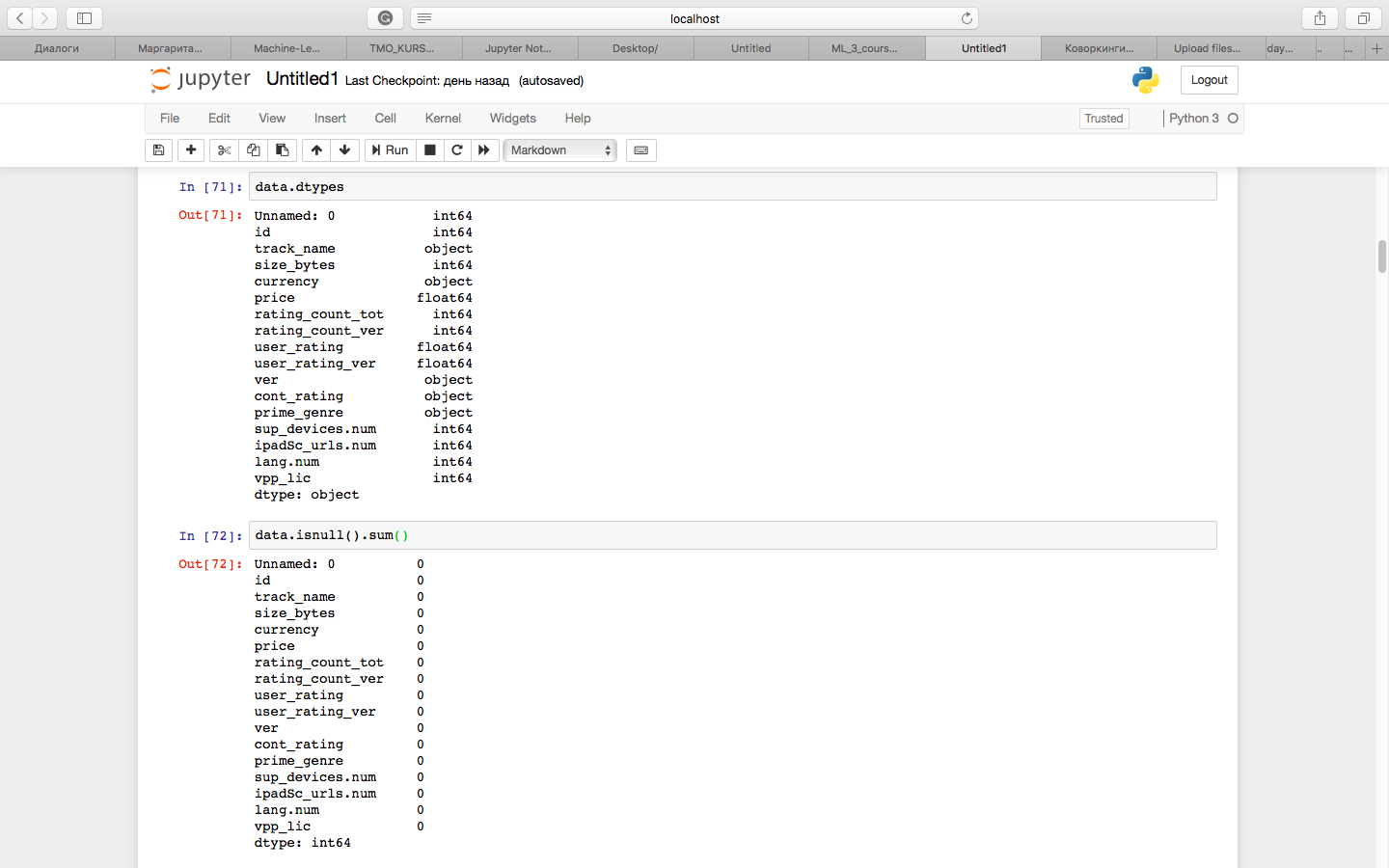
"vpp\_lic": лицензирование обработки векторных пакетов на основе устройств включено

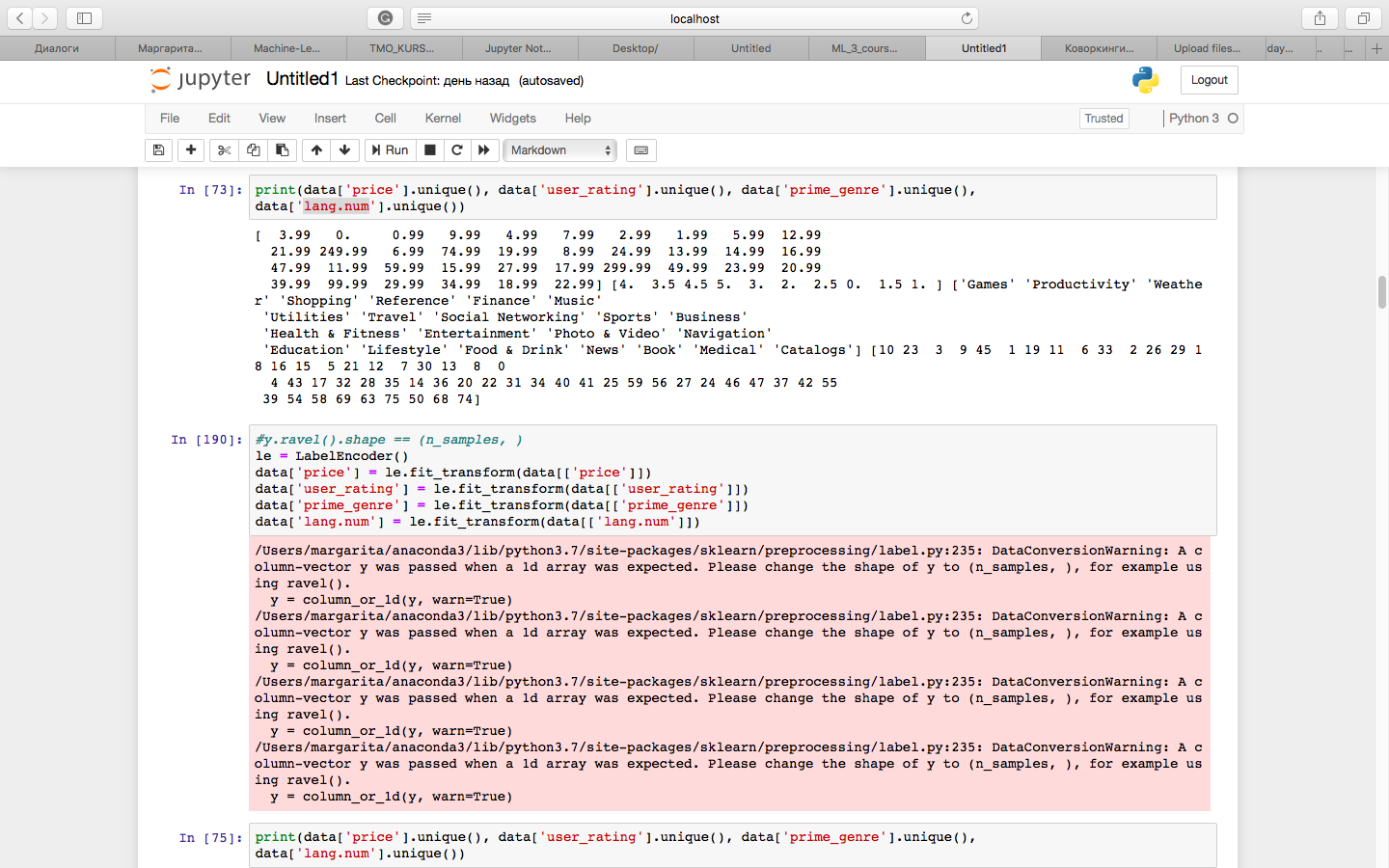
"prime\_genre": основной жанр. Я выбрала его целевым признаком.

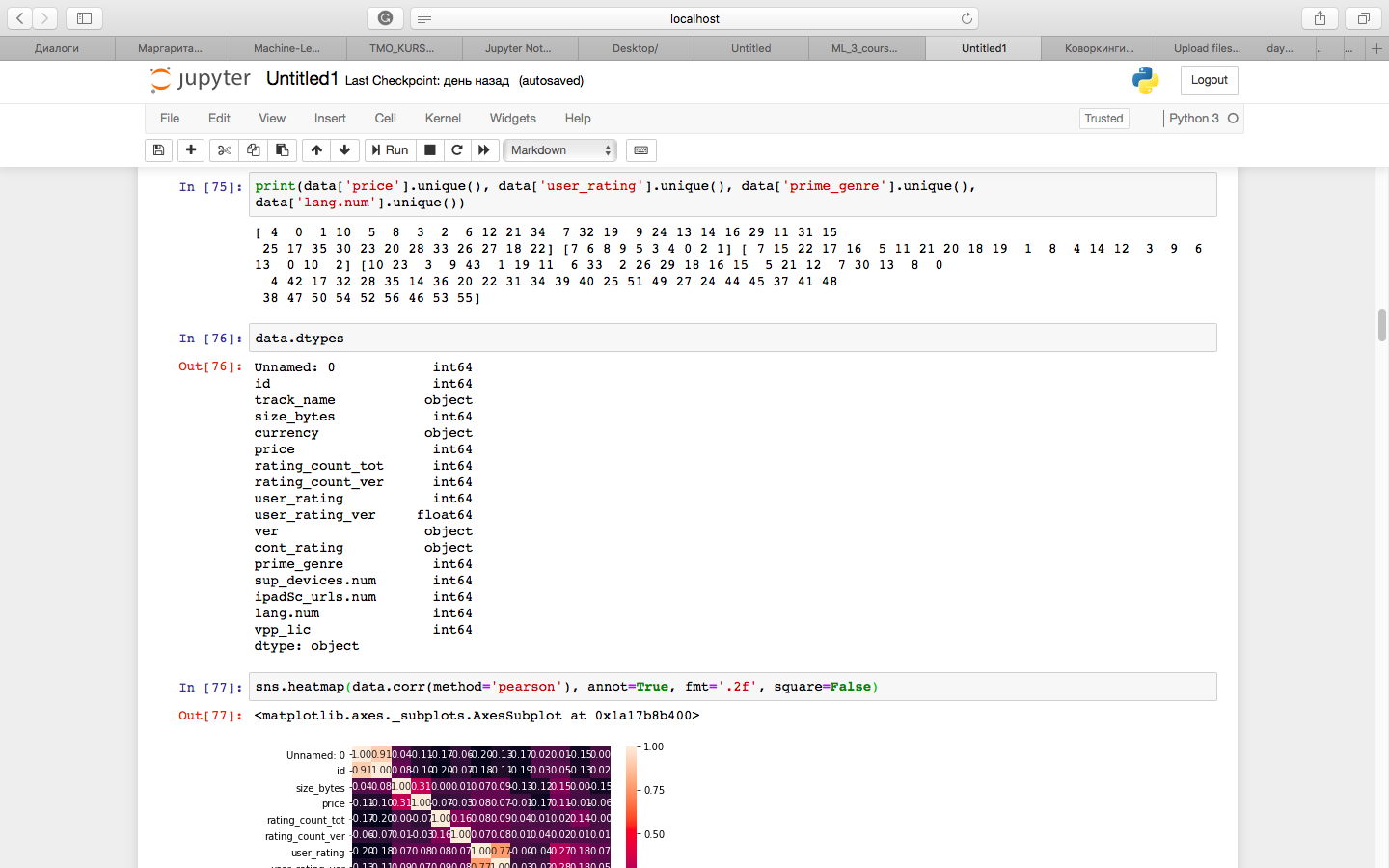
**Текст программы с примерами выполнения**

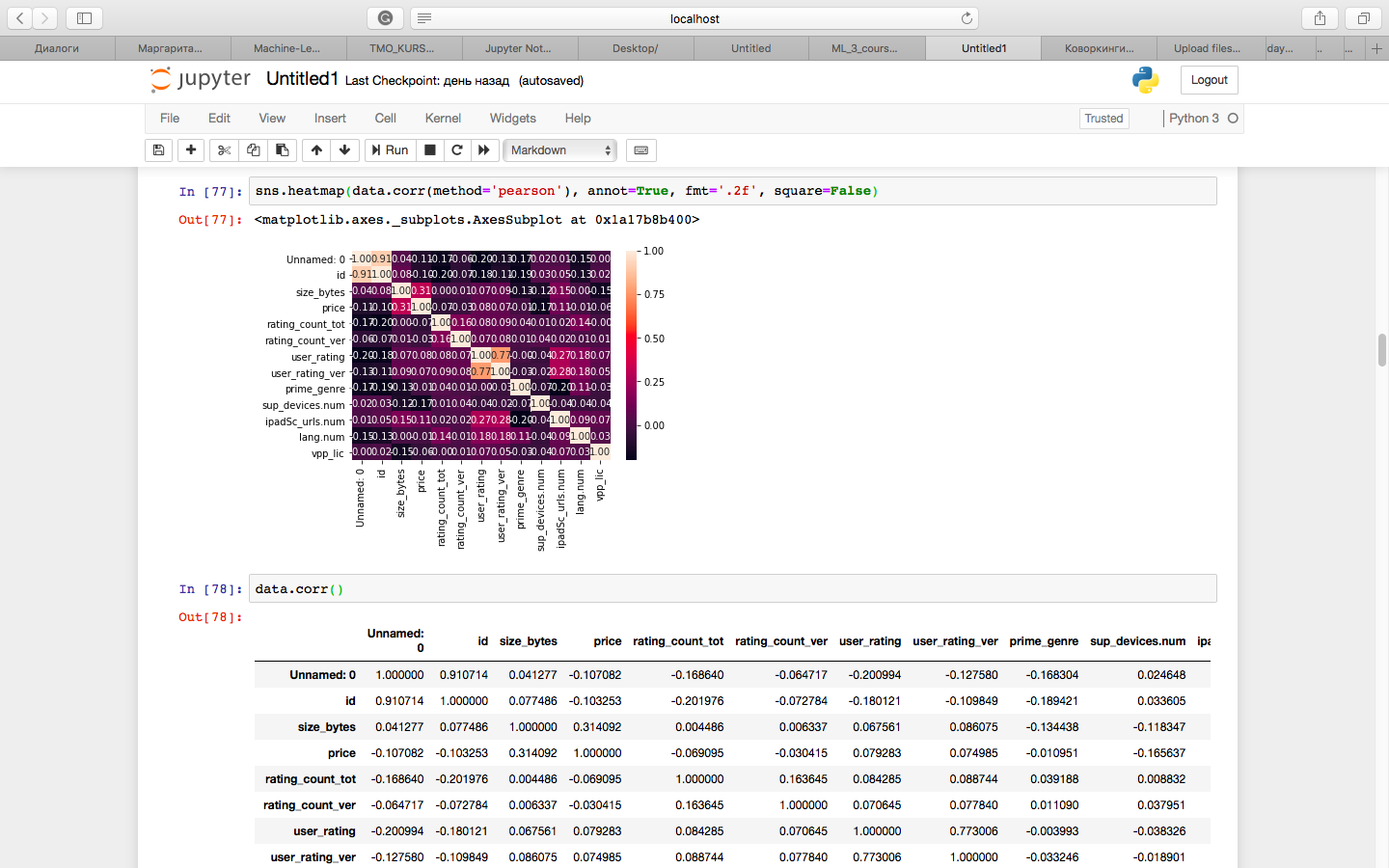


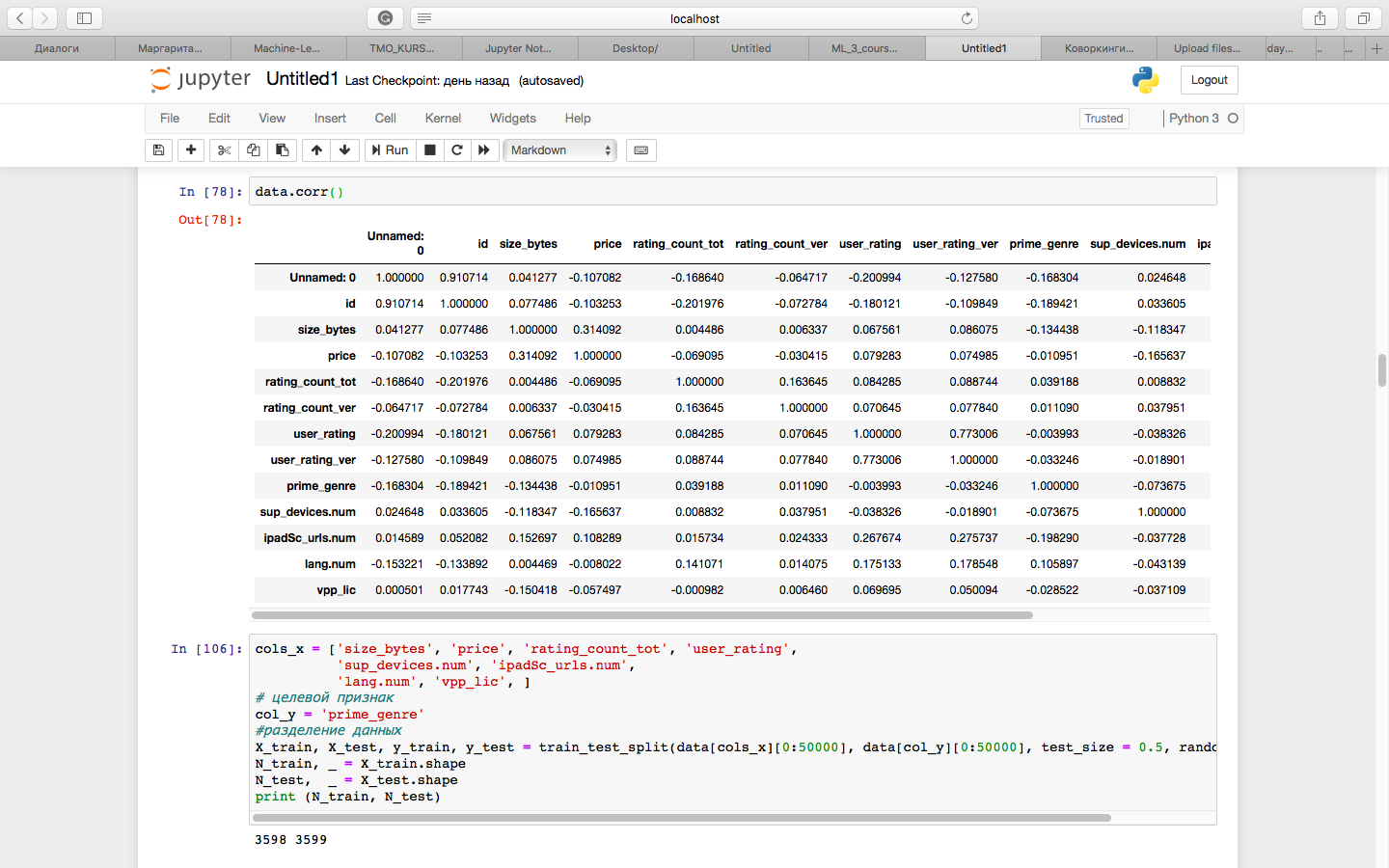




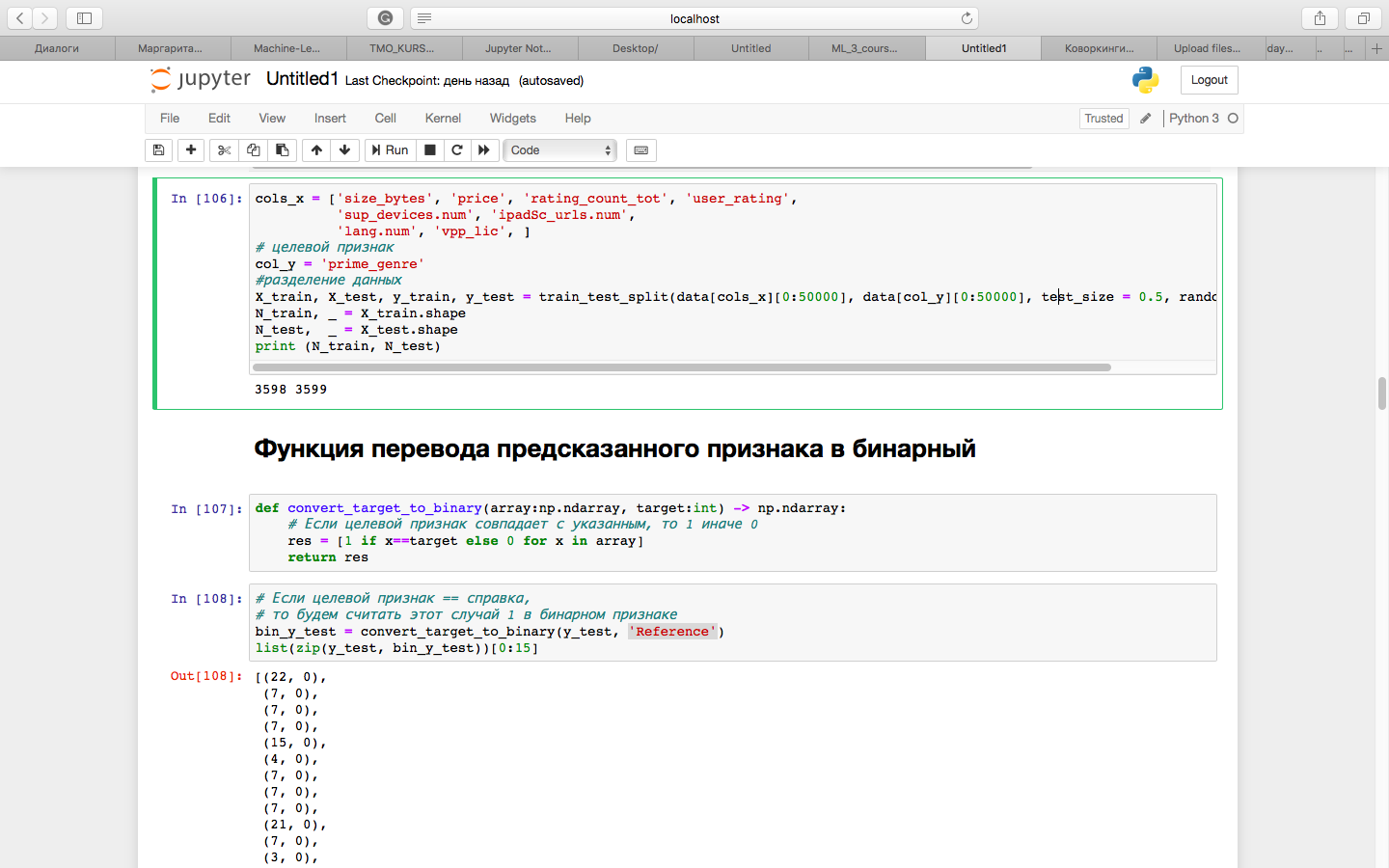


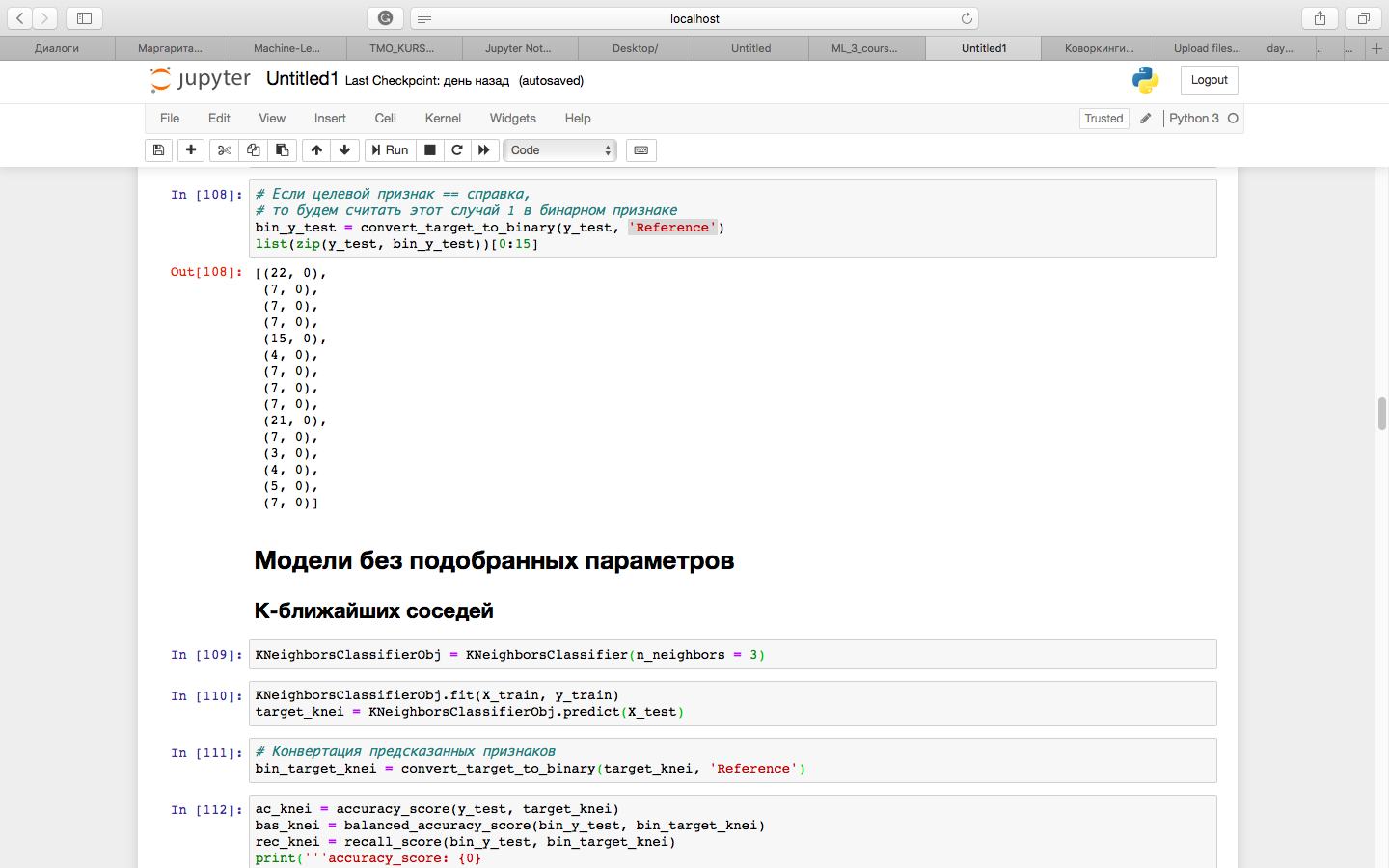


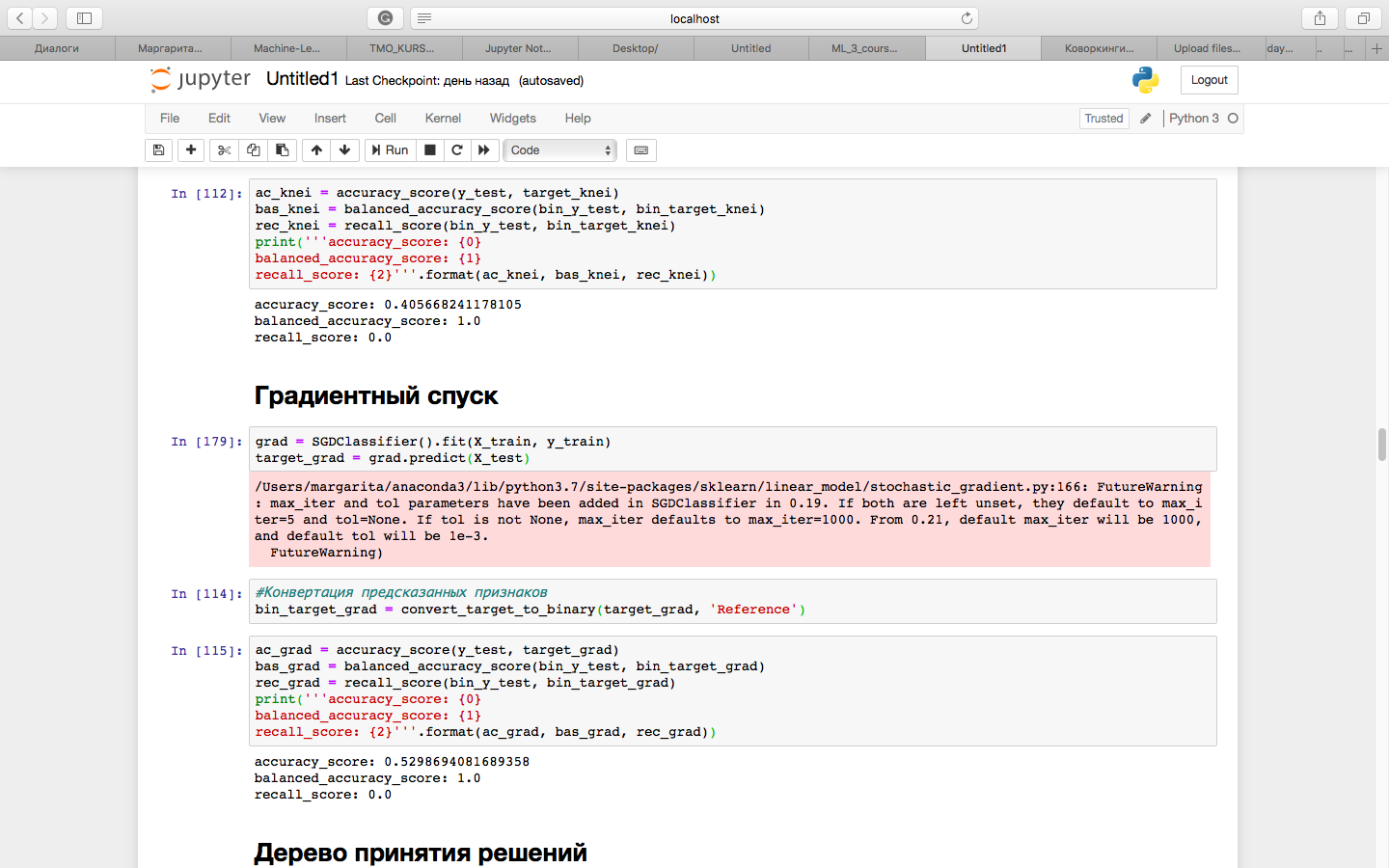


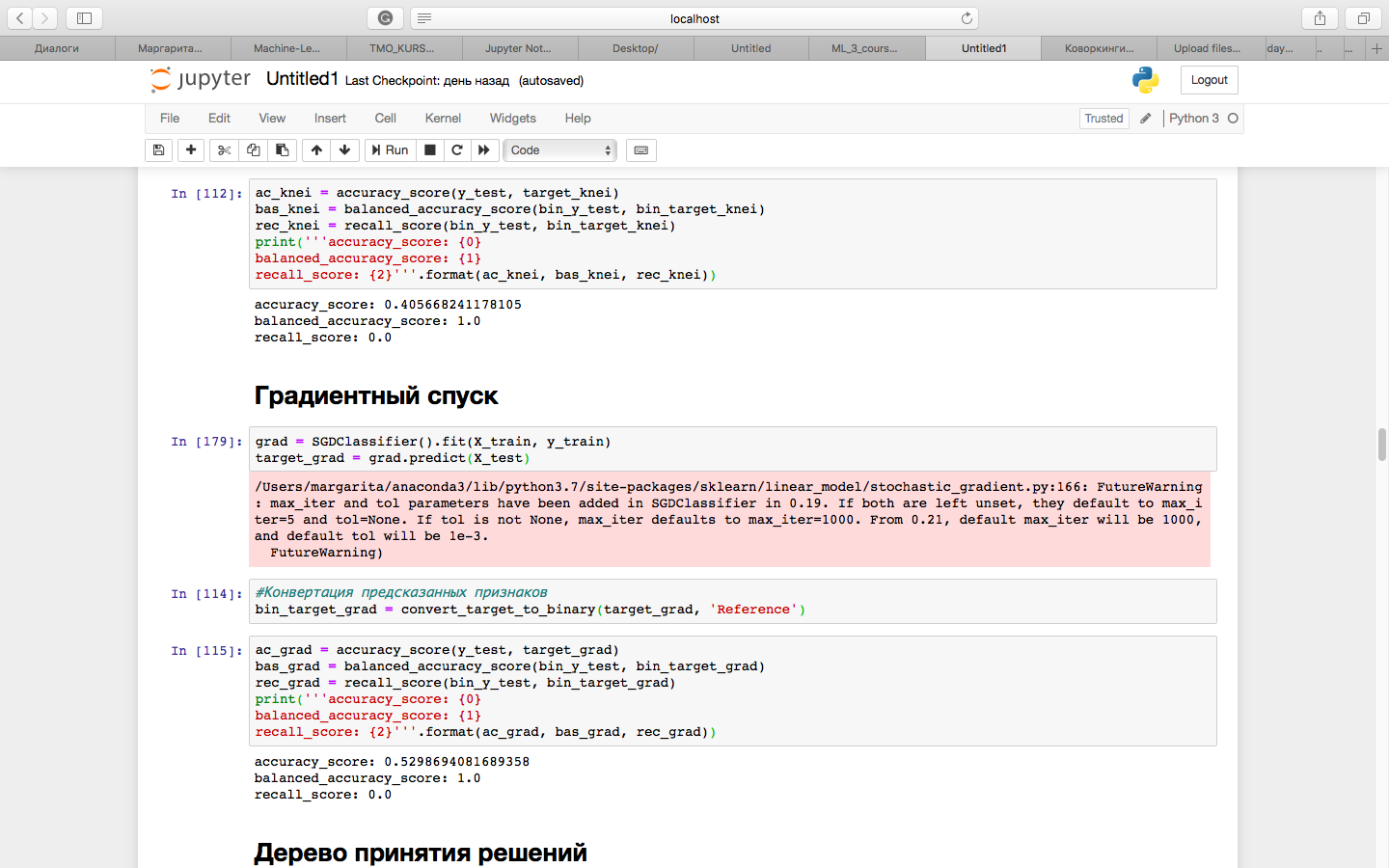


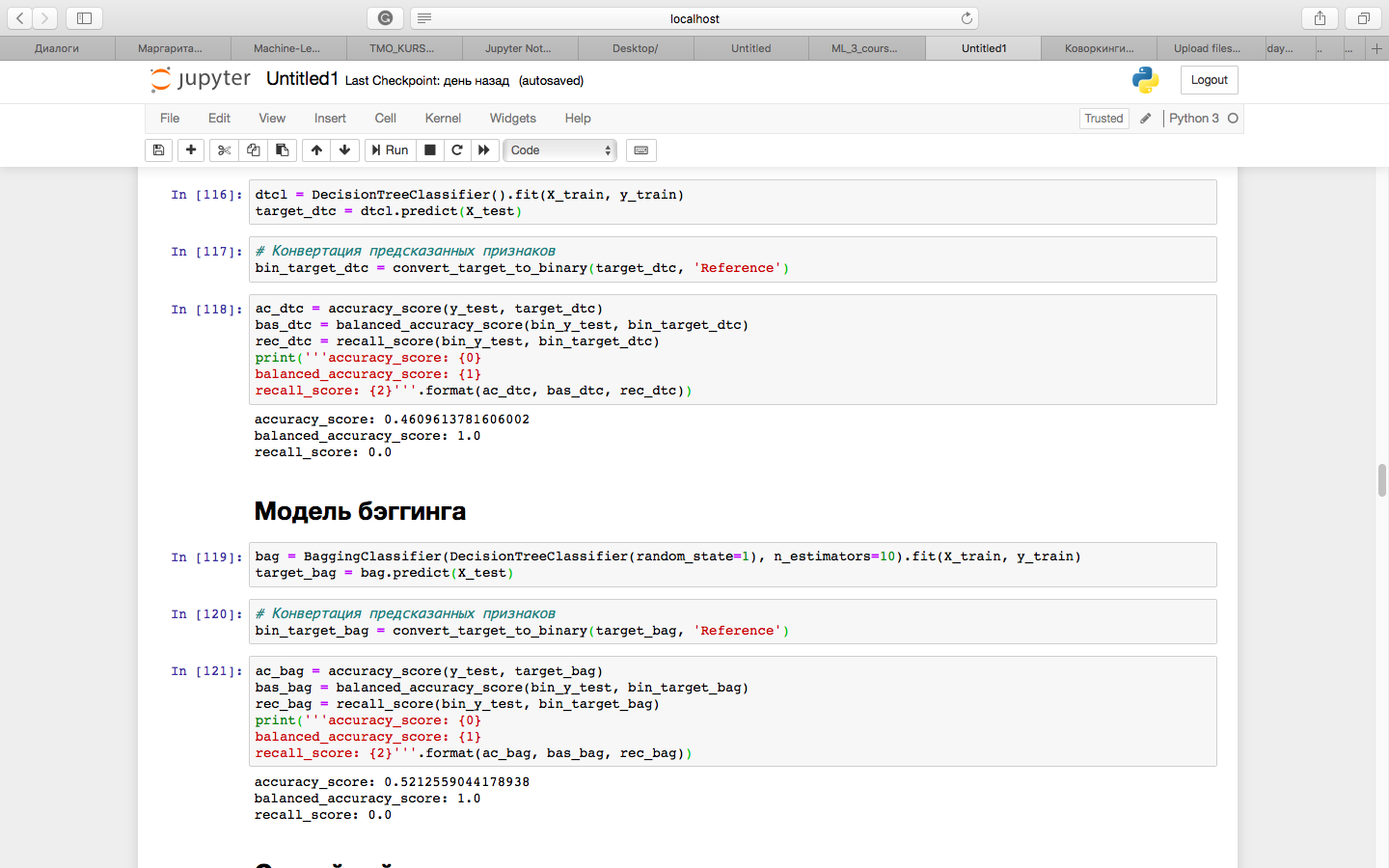
Мной быливыбраны следующие фичи для последующей работы с ними:

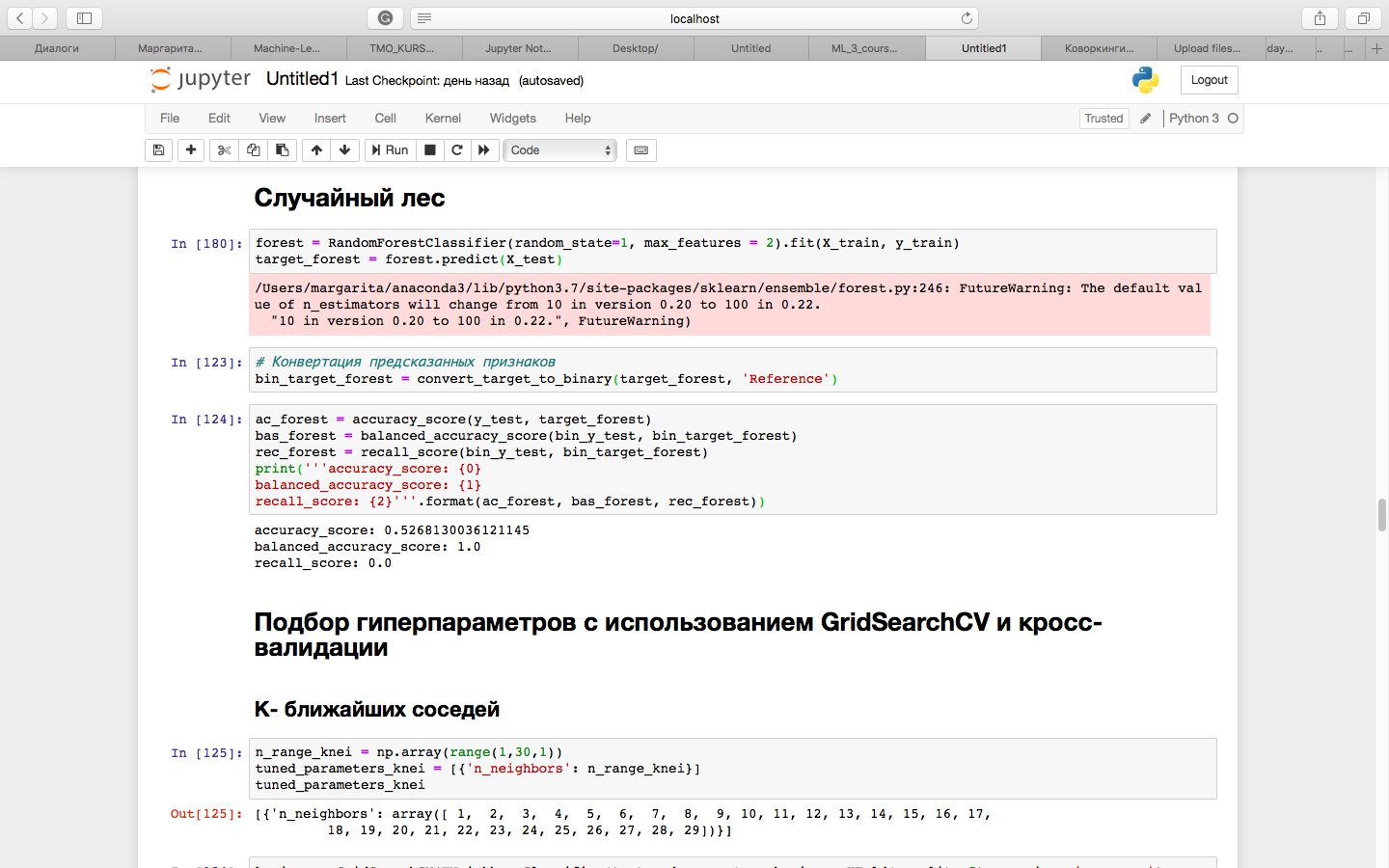


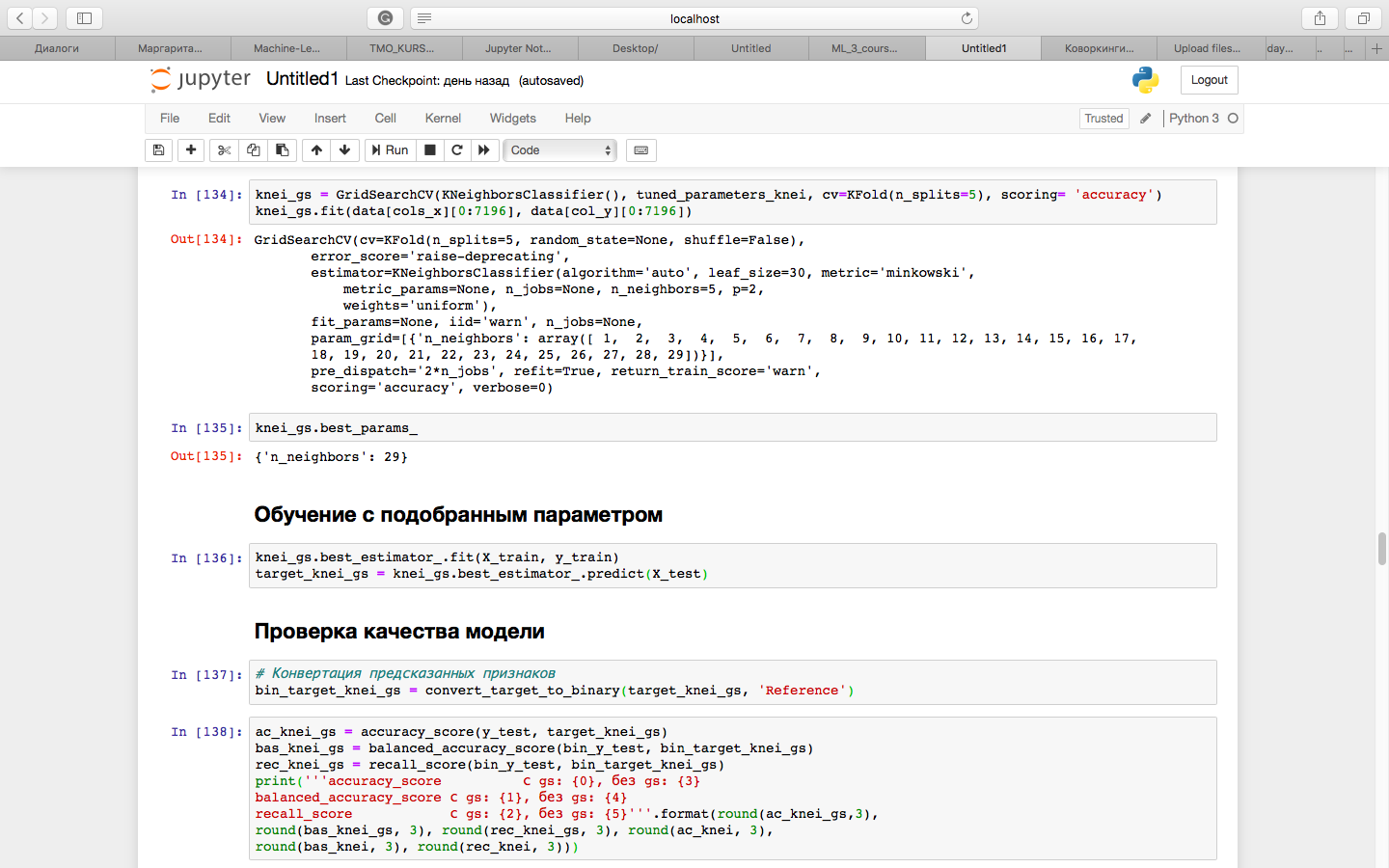


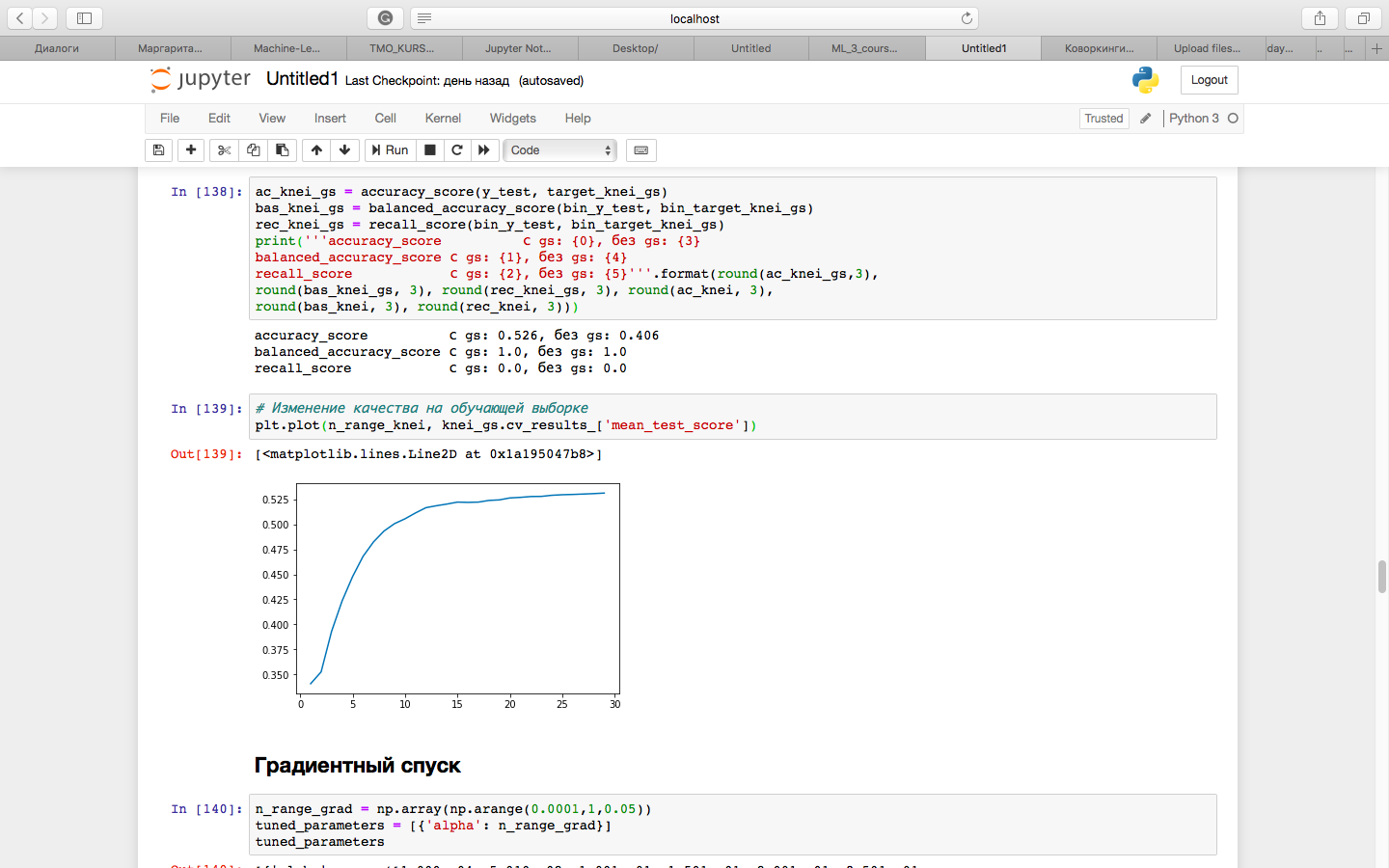


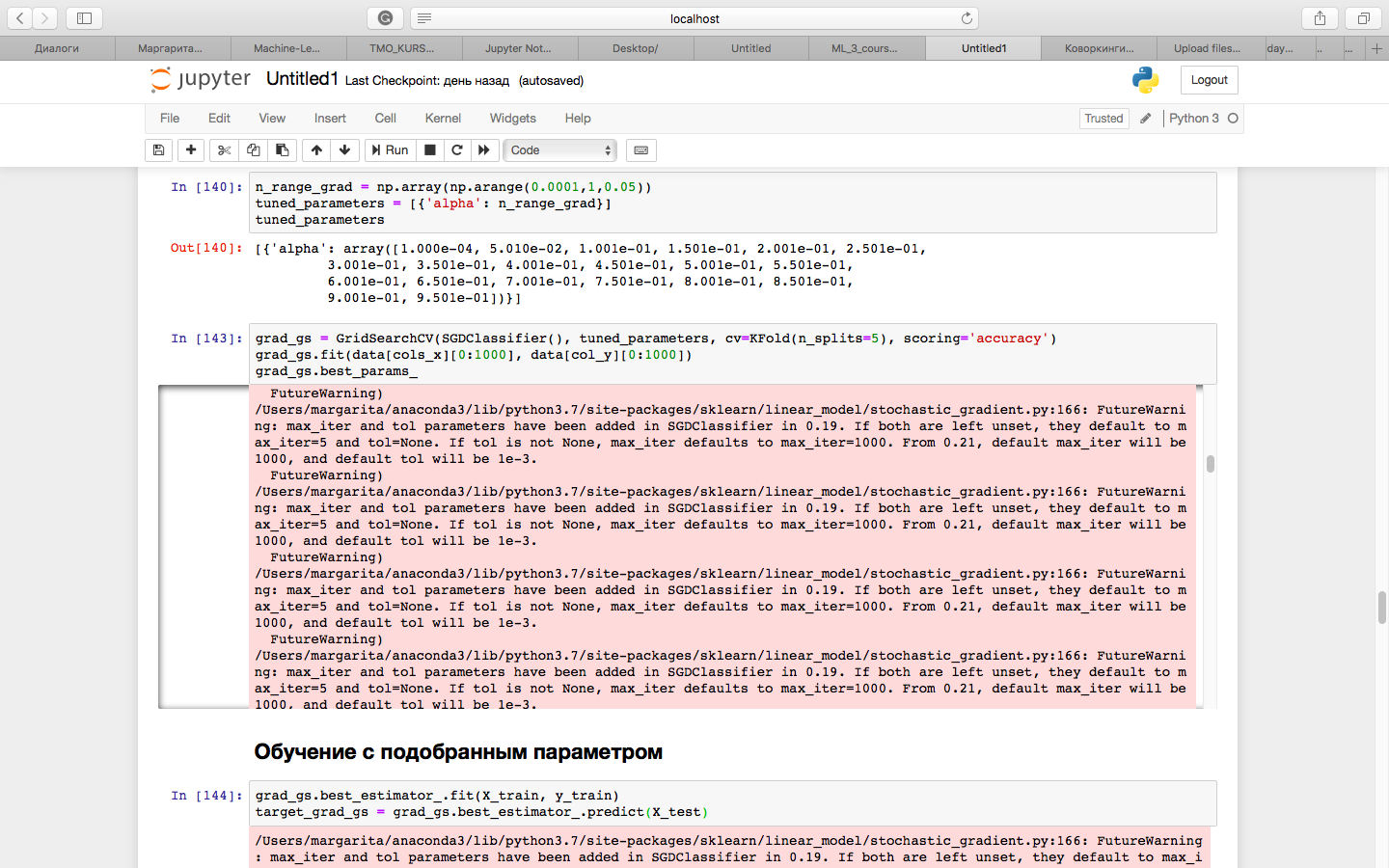


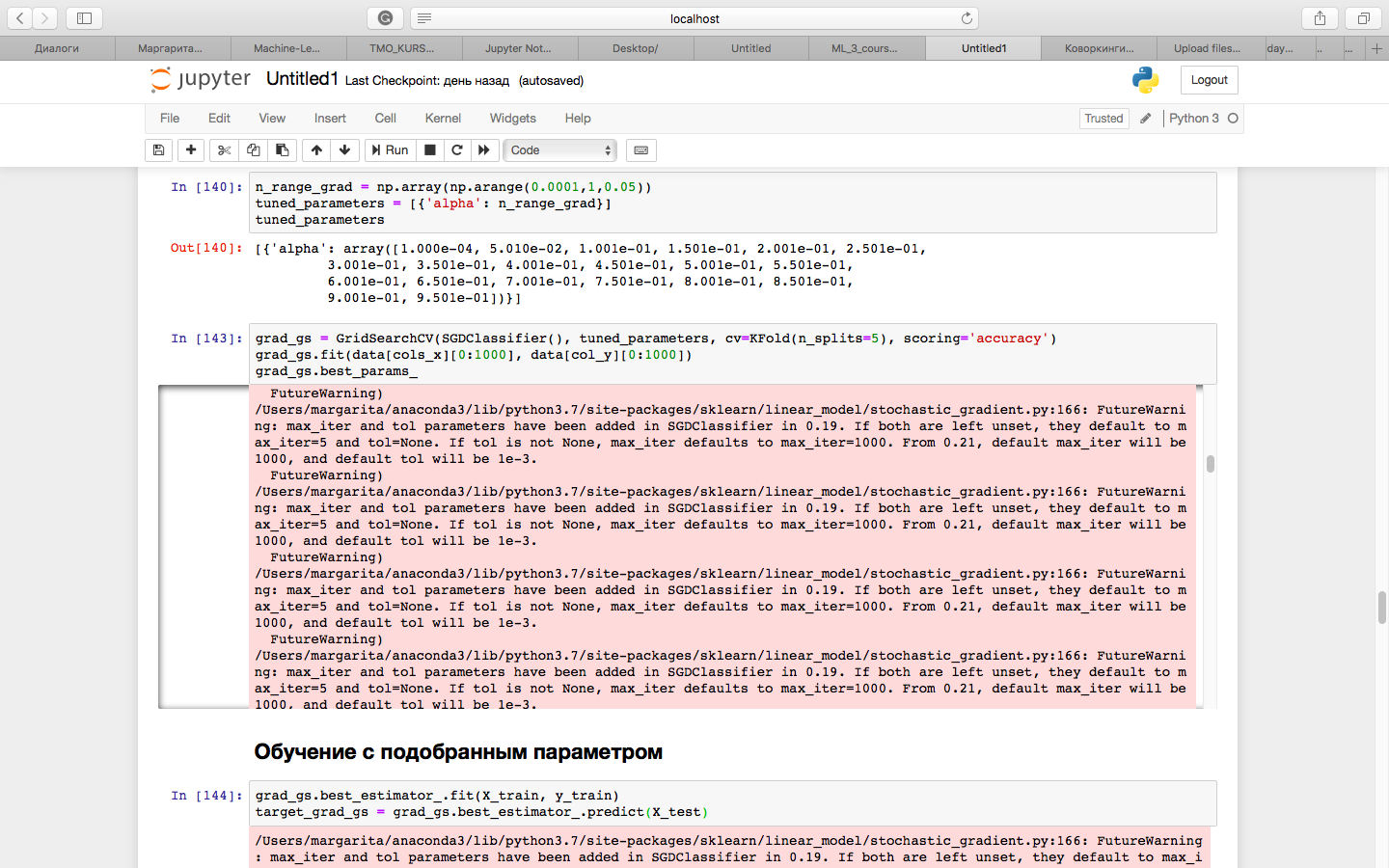


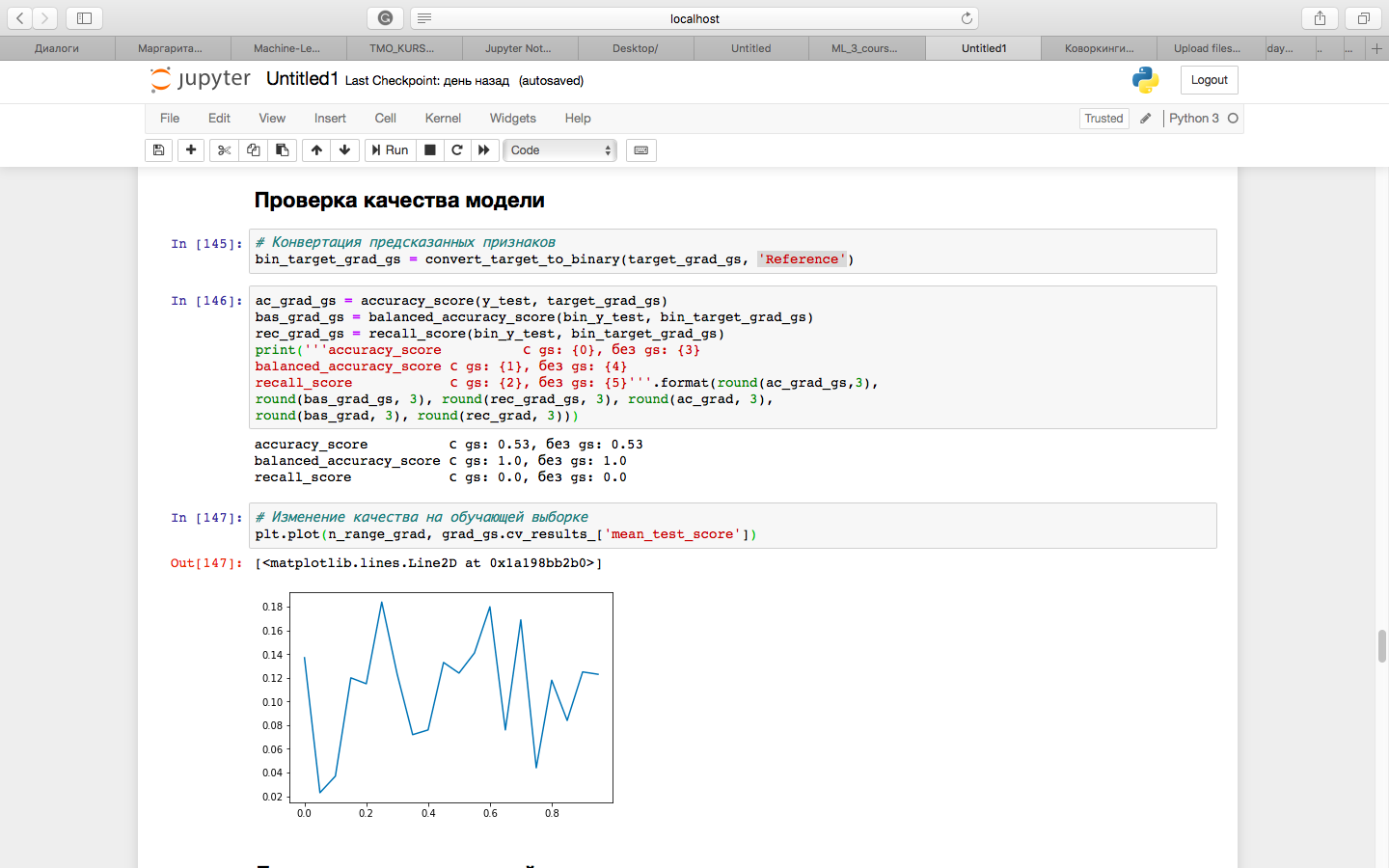


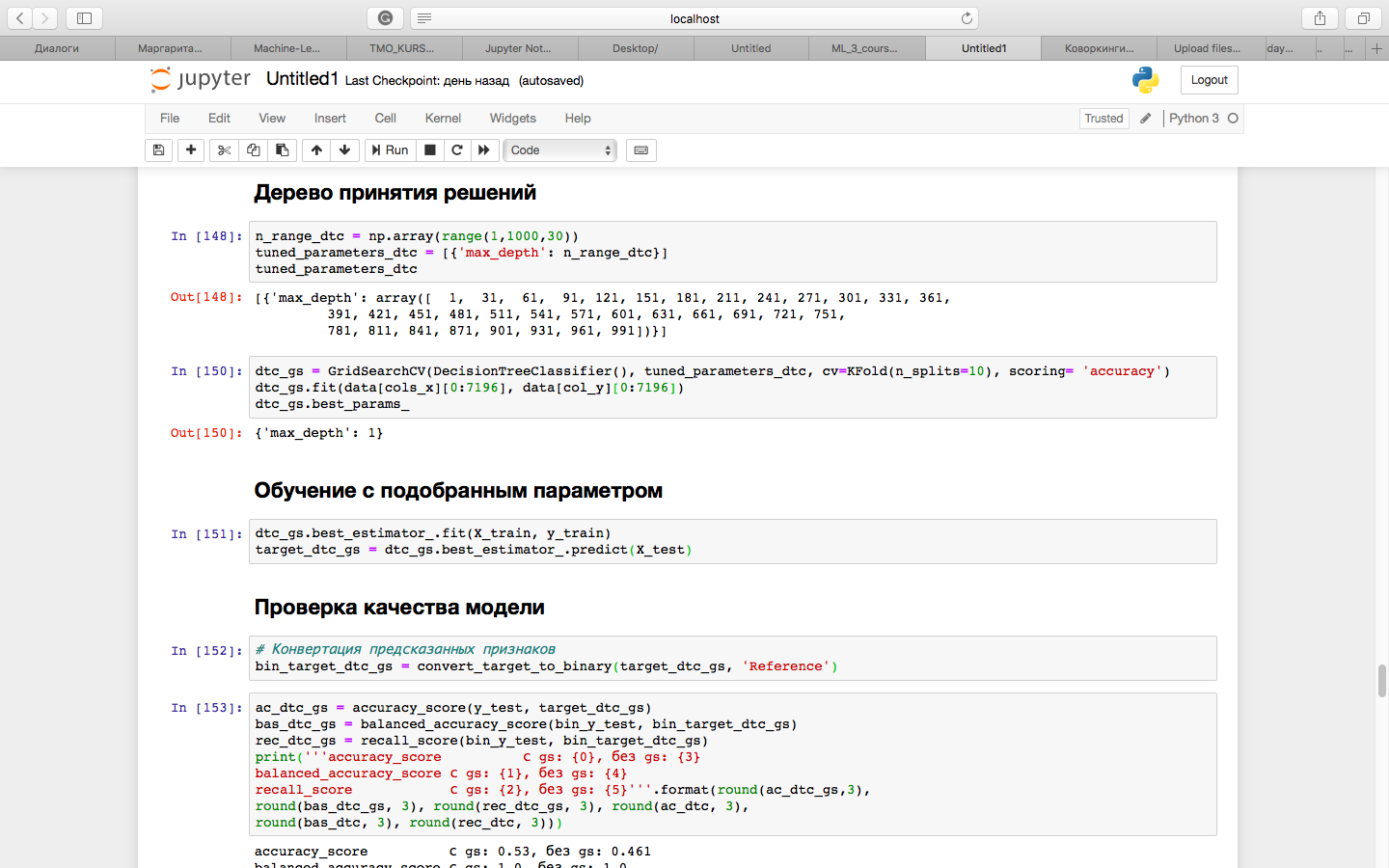




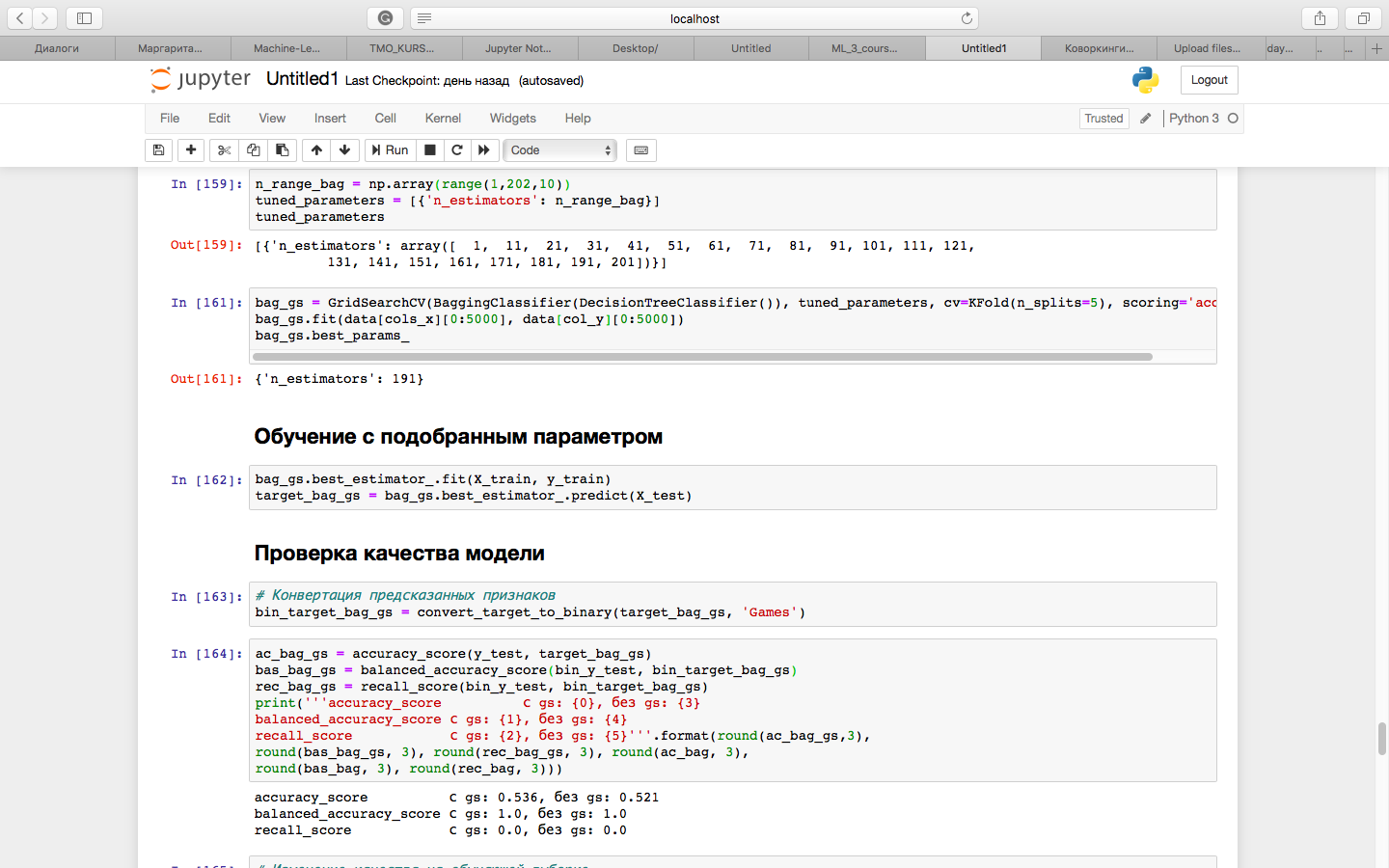


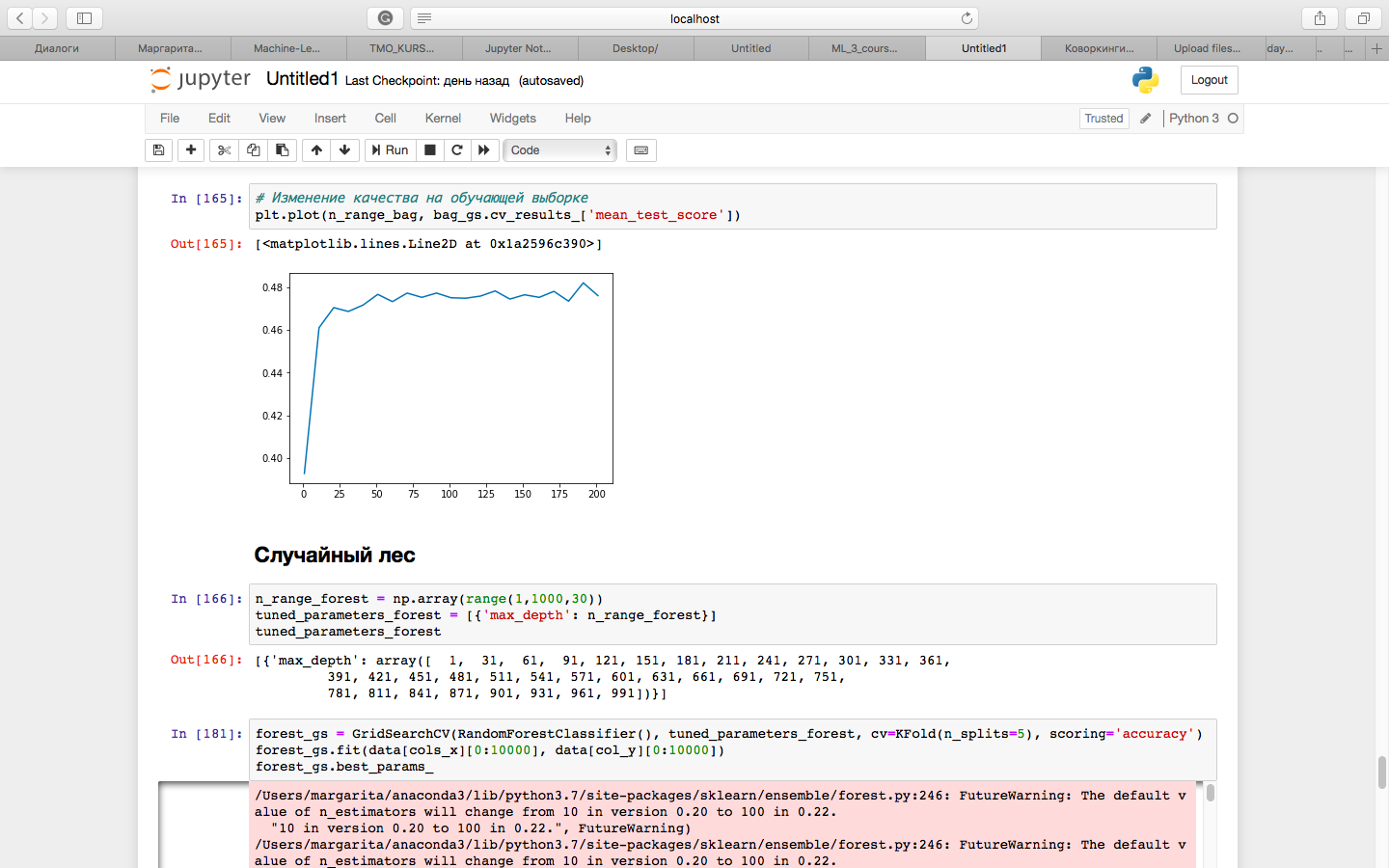


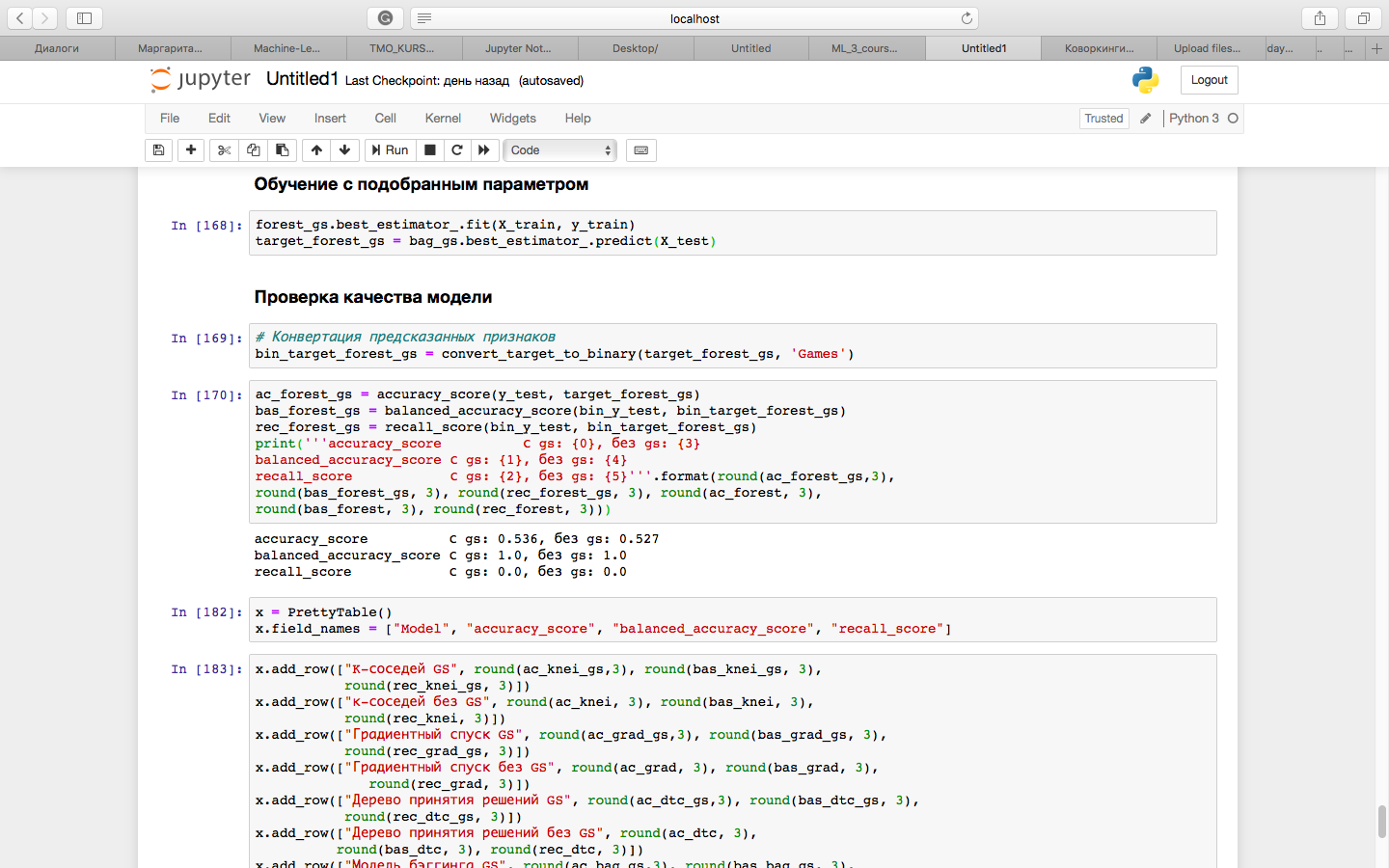


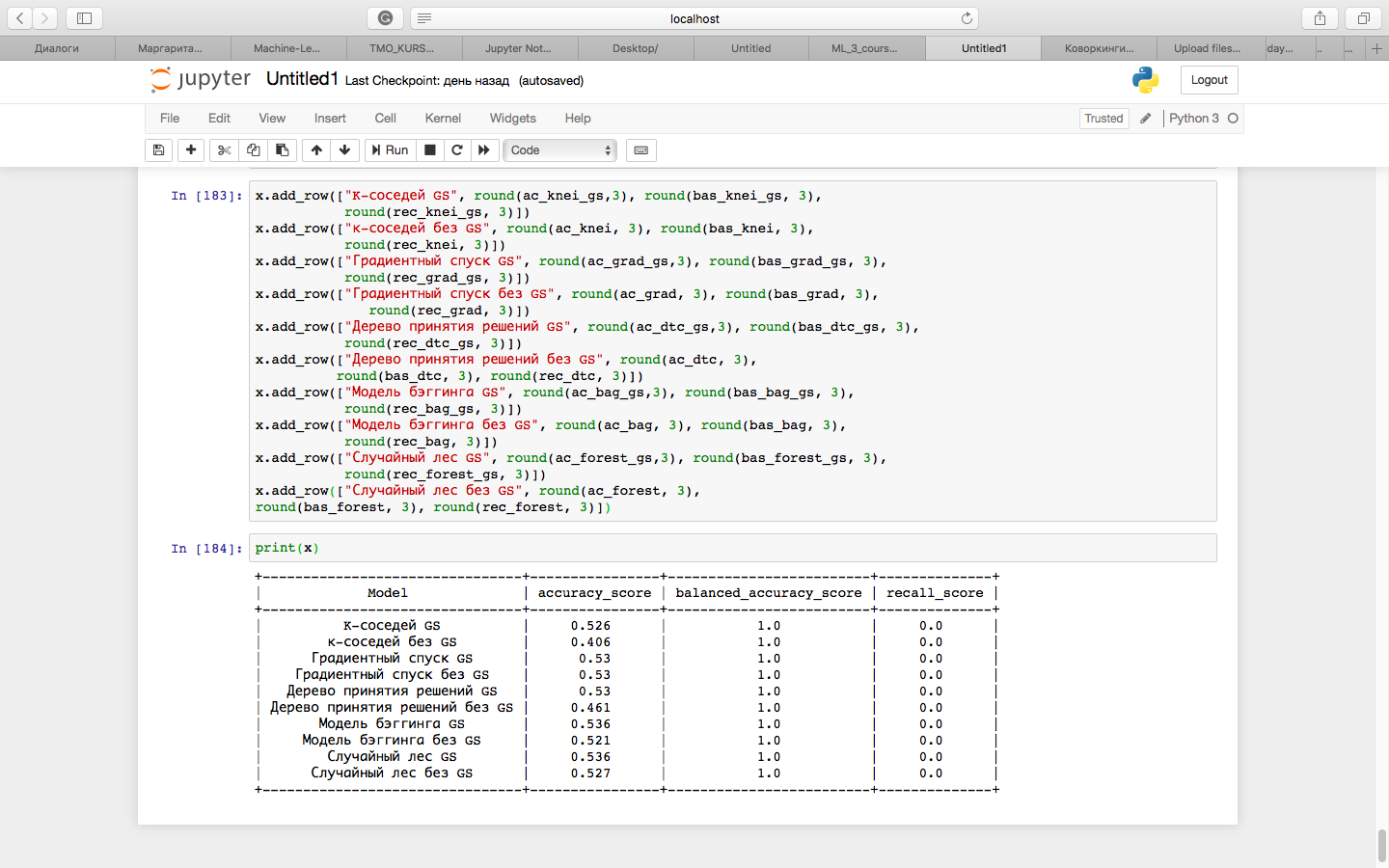












**Заключение**

Таким образом, с помощью машинного обучения можно направленно рекламировать функции своего продукта и получать от этого максимальный прирост пользователей различных функций. Наилучший результат показала модель бэггинга. В результате некоторых доработок данная модель может быть вполне использована на практике.

**Список литературы**

1. Лекции Гапанюка Ю. Е. [Электронный ресурс]. – Электрон. дан. – URL: <https://github.com/ugapanyuk/ml_corse/wiki/CORSE_TMO>
2. Mobile App Statistics (Apple iOS app store) Data Set [Электронный ресурс]. – Электрон. дан. – URL: <https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps/downloads/app-store-apple-data-set-10k-apps.zip/7>
3. Stackoverflow [Электронный ресурс]. – Электрон. дан. – URL: https://stackoverflow.com
4. Scikit-learn [Электронный ресурс]. – Электрон. дан. – URL: https://scikit-learn.org/stable/#