г.Москва, 2019

Строков Никита

М80-306Б

Лабораторная работа №1

По курсу машинного обучения

# Московский Авиационный Институт (Национальный исследовательский университет)

# Факультет «Информационные технологии и прикладная математика»

# Кафедра 806 «Вычислительная математика и программирование»

Постановка задачи

Познакомиться с платформой Azure Machine Learning, реализовывав полный цикл разработки решения задачи машинного обучения, использовав три различных алгоритма, реализованные на этой платформе.

Требования:

1. Уникальность решения
2. Обоснованность выбора той или иной операции
3. В отчете должны быть указаны алгоритмы, которые применялись, результаты применения этих алгоритмов, а также скрины некоторых этапов обработки данных

Выполнение работы

## Предсказание численного значения

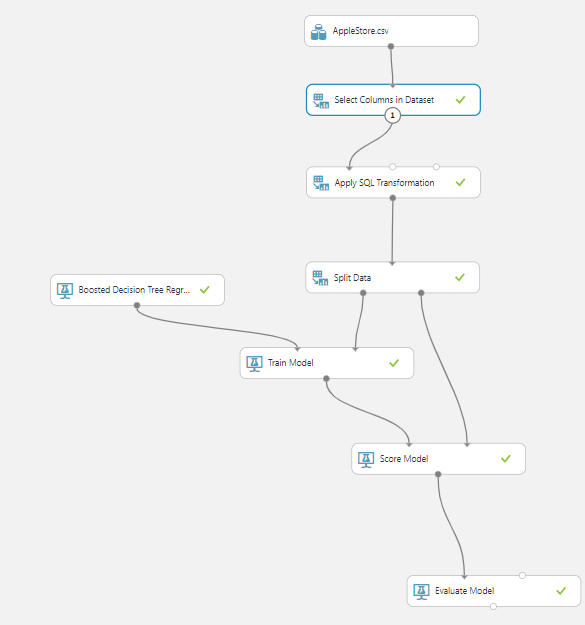
Будем использовать регрессию для определения ожидаемого рейтинга версии.

Предсказываемый атрибут:

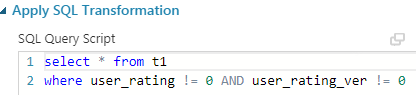
* user\_rating\_ver

Атрибуты для предиктивной модели:

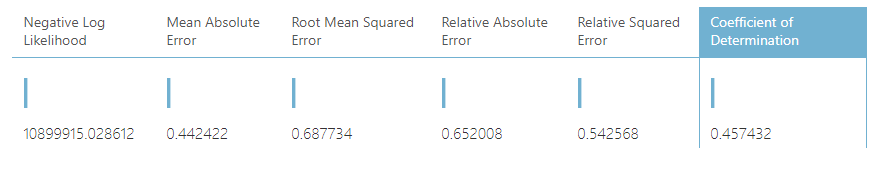
* price
* rating\_count\_tot
* user\_rating
* sup\_devices.num
* lang.num



В данной модели для обучения SQL запросом отсекаюстя те ряды, в которых значение user\_rating или user\_rating\_ver равно нулю.



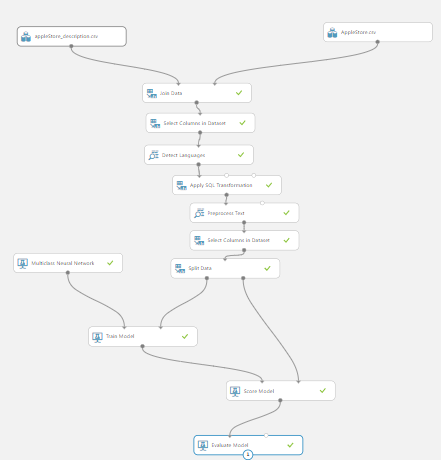
По сравнению с другими алогритмами регрессии, алгоритм Boosted Decision Tree предсказывал наиболее точно значения атрибута, поэтому используется именно он.



Шкала Negative Log Likehood говорит о том, что данная модель предсказывания не очень точная, но по ней можно предугадывать, но точную оценку этой версии, а пойдет ли оценка вверх, вниз или останется примерно на таком же уровне.

## Предсказывание категориальных признаков

В этой предиктивной модели будем предугадывать жанр приложения по описанию.

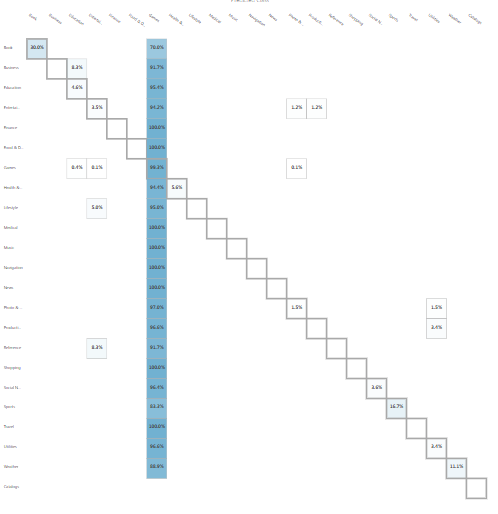


Так как в датасете с текстом нет атрибута prime\_genre, то объединим датасеты по полю id, а затем выберем колонны *prime\_genre, size\_bytes, app\_desc*.

В датасете присутствуют описания приложений не только на английском, но и на японском и китайском. Чтобы избавиться от них, воспользуемся функцией Detect Languages, а затем SQL запросом оставим только ряды с английским языком.

Затем сделаем обработку текста, которая «очистит» наш текст. Далее для предсказывания категориальных признаков используем алгоритм Multiclass Neural Network и обучим модель.

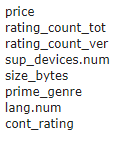
Получили данную оценку точности модели

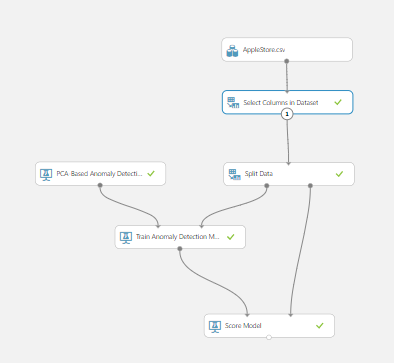


Как мы видим, точность в данной ситуации плохая. Это вызвано тем, что приложений жанра “Games” было минимум в 7 раз больше, чем в каждом из жанров. Возможно, имело смысл сделать двухклассовую предиктивную модель, которая предсказывала бы является приложение игрой или нет.

Поиск аномалий.

Будем искать аномалии среди следующих атрибутов:





Вывод

Работая над данной лабораторной работой я, познакомился с платформой Azure Machine Learning, реализовывал полный цикл разработки решения задачи машинного обучения, использовав три различных алгоритма, реализованные на этой платформе.