

# 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, реализовав 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

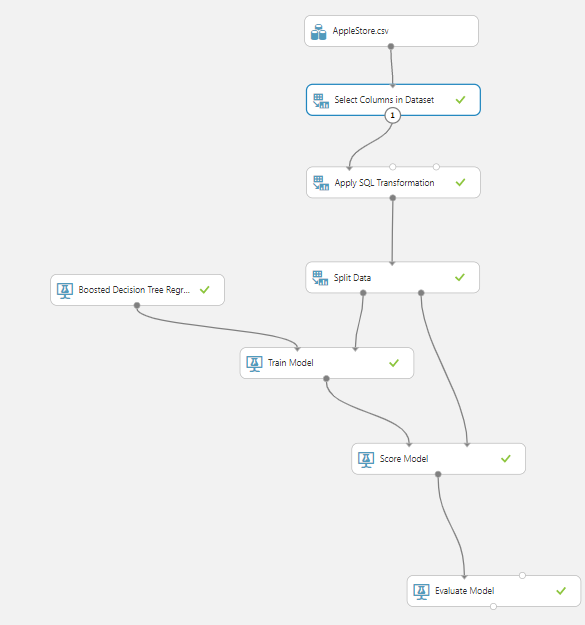
Будем использовать регрессию для определения ожидаемого рейтинга версии.

Предсказываемый атрибут:

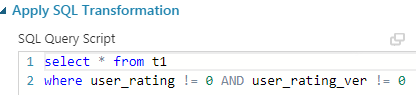
* user\_rating\_ver

Атрибуты для предиктивной модели:

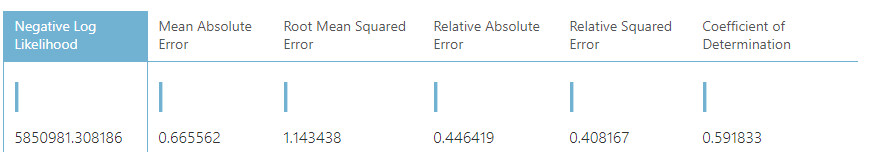
* price
* rating\_count\_tot
* user\_rating
* sup\_devices.num
* lang.num
* cont\_rating
* prime\_genre
* ipadDc\_urls.num
* size\_bytes



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

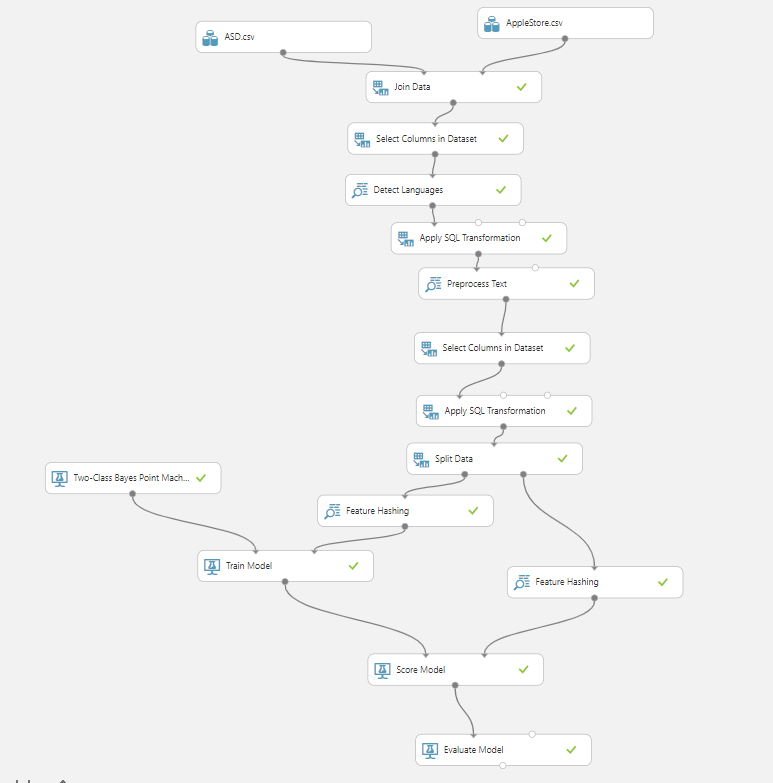


По сравнению с другими алогритмами регрессии, алгоритм Boosted Decision Tree предсказывал наиболее точно значения атрибута, поэтому используется именно он.



## Предсказывание категориальных признаков

В этой предиктивной модели будем предугадывать жанр приложения по описанию.



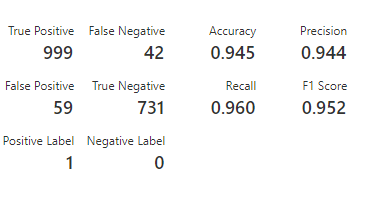
Так как в датасете с текстом нет атрибута prime\_genre, то объединим датасеты по полю id, а затем выберем колонны *prime\_genre, app\_desc*.

Перед работой с датасетом, пришлось убрать все newline'ы и кавычки, чтобы ML Studio полностью считывала текст.

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.

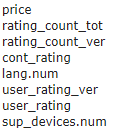
Затем сделаем обработку текста, которая «очистит» наш текст. Далее для двухклассовой классификации используем Наивный Байесовский Классификатор и хеширование.

Получили данную оценку точности модели

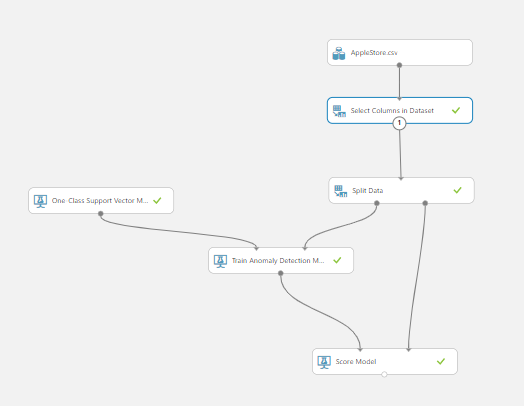


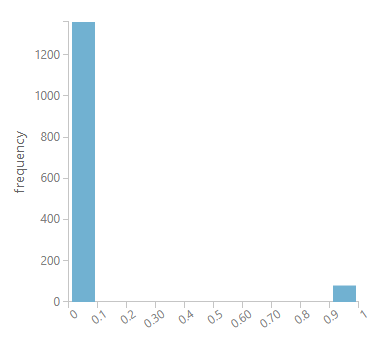
## Поиск аномалий.

Будем искать аномалии среди следующих атрибутов:



Используется метод *One-Class Support Vector Machine.*





В итоге модель нашла 5.6 % аномальных записей.

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.