**Московский авиационный институт**

**(Национальный исследовательский университет)**

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Искусственный интеллект»

**Лабораторная работа № 1**

Тема: Знакомство с Azure Machine Learning

Студент: Чекушкин Д.И

Группа: М80-304Б

Преподаватель: Ахмед Самир Халид

Москва, 2019

1. Постановка задачи

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

1. Решения задачи

Датесеты взяты из ЛР0:

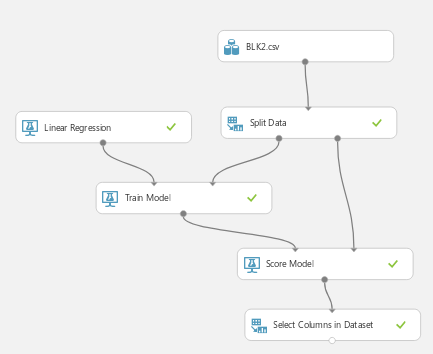
1) <https://finance.yahoo.com/quote/BLK/history?period1=941403600&period2=1553288400&interval=1d&filter=history&frequency=1d&guccounter=1> История акций компании BlackRock за все время существования

2) <https://www.kaggle.com/snapcrack/all-the-news> Новости, взятые из разных источников

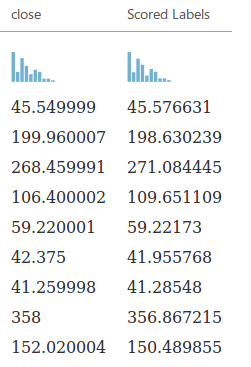
**Алгоритм 1** - оценка стоимости закрытия акций с помощью регрессии.

Будем использовать линейную регрессию, так как в прошлой ЛР удалось обнаружить линейную зависимость между атрибутами акций.

На основе имеющихся данных (дата, цена открытия, цена закрытия, max цена, min цена, скорректированная цена, объем) обучим модель определять цену закрытия. Соотношения данных на обучение/проверку = 80/20.

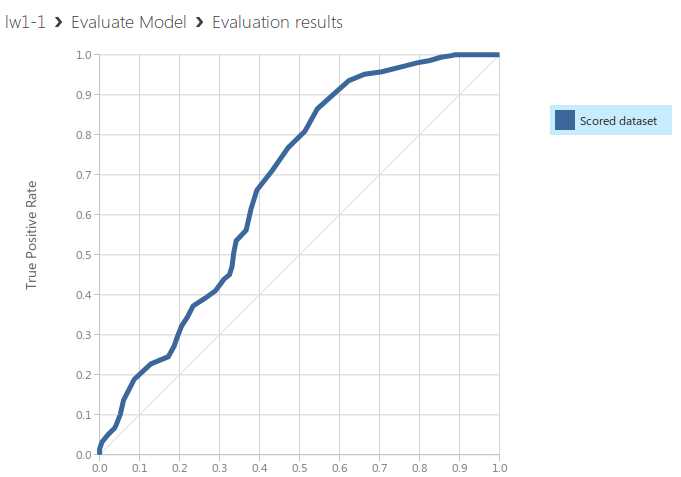


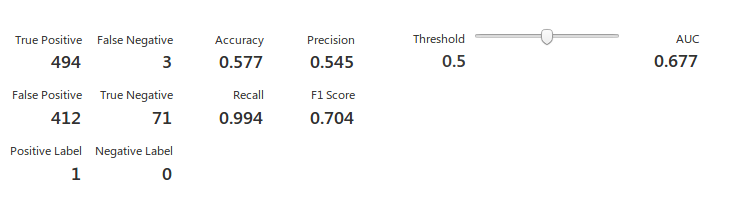
И результат:



Получилось довольно близко.

Если добавить столбец result, состоящий из 1 и 0(есть прибыль или нет), то можно воспользоваться two-class logistic regression алгоритмом и получить:



Что мы видим?

494 True positive - результат классификации положительный, и истинное значение тоже положительное

412 False positive - результат классификации положительный, но истинное значение отрицательное

3 False-negative - результат классификации отрицательный, но истинное значение положительно

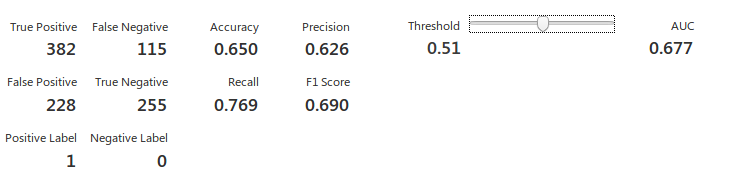
71 True-negative - результат классификации отрицательный, и истинное значение тоже отрицательное

Precission = tp/(tp+fp) - мера точности правильных ответов. Чем больше точность, тем меньше число ложных предсказываний.

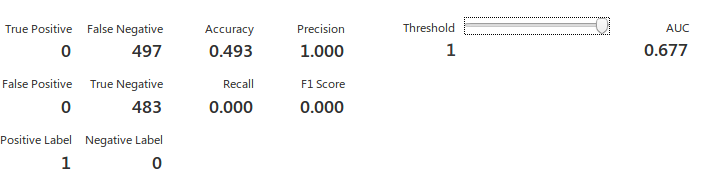
Recall = tp/(tp+fn) - мера полноты (способность классификатора «угадывать» как можно большее число положительных ответов из ожидаемых)

Значение коэффициента threshold - порог, значения выше которого указывает на наличие прибыли(1), а ниже на отсутствие (0).

При увеличении значения порога: false-positive значений стало значительно меньше. , но при этом возросло количество false-negative(классификатор их не пропустил). Recall уменьшился, а Precission вырос.

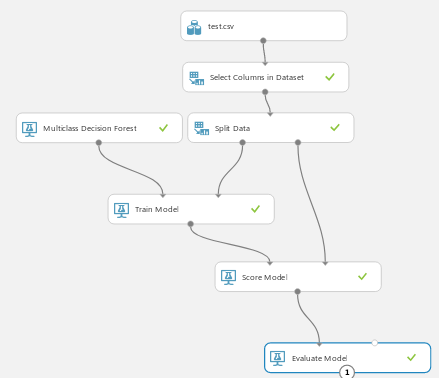


Если увеличить Threshold до максимума, то получим

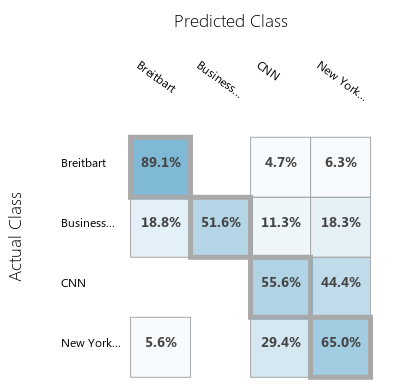


Сразу обратим внимание на Recall = 0. Все правильные и неправильны positive значения не отобраны и поэтому Presicion равен 1.

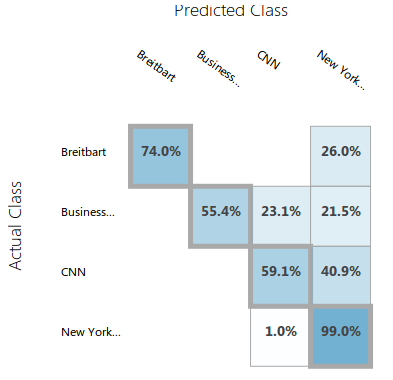
**Алгоритм 2** - Классификация новостных статей по издателю из датасета 2. Для этого воспользуемся мультиклассовым алгоритмом(так как 4 издателя).



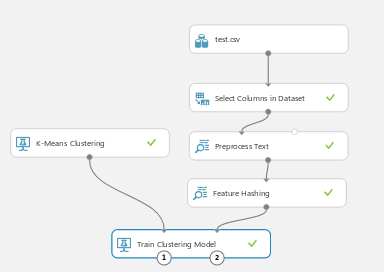
Результат:



На этот раз не все так идеально. Два издателя предсказываются с довольно высокой точностью 89 и 65%, оставшиеся два менее точно. Если убрать из датасета дату публикации, то результаты становятся точнее



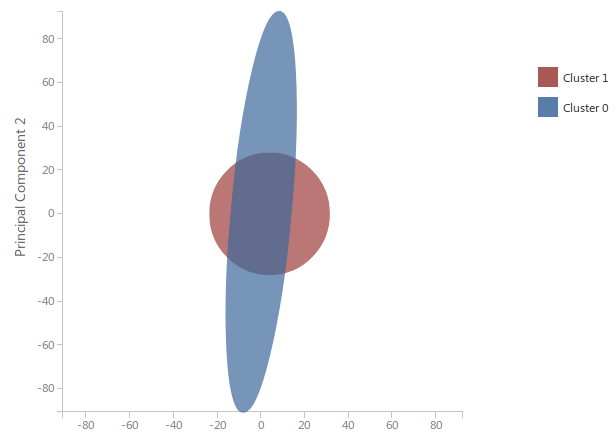
**Алгоритм 3** — кластеризация заголовков и издателей.



Для начала выберем интересующие нас колонки — издатель и статья.

Далее обратимся к встроенной функции — Preprocess text (чтобы избавиться от «шума» в данных) и Feature Hashing, для преобразования потока текста на английском языке в набор функций, представленных в виде целых чисел.

Получаем результат:



Два кластера данных.

Вывод:

Благодаря проделанной работе, мне удалось познакомиться с платформой Azure Machine Learning и реализовать три различных алгоритма. Я научился прогнозировать, классифицировать и кластеризировать данные. Очень хотелось бы продолжить работать с этой платформой.