

Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной
математики
Кафедра вычислительной математики и программирования

Лабораторная работа № 1 по курсу "Искусственный интеллект"

Студент: А. В. Скворцов
Группа: М8О-308Б

Москва, 2019

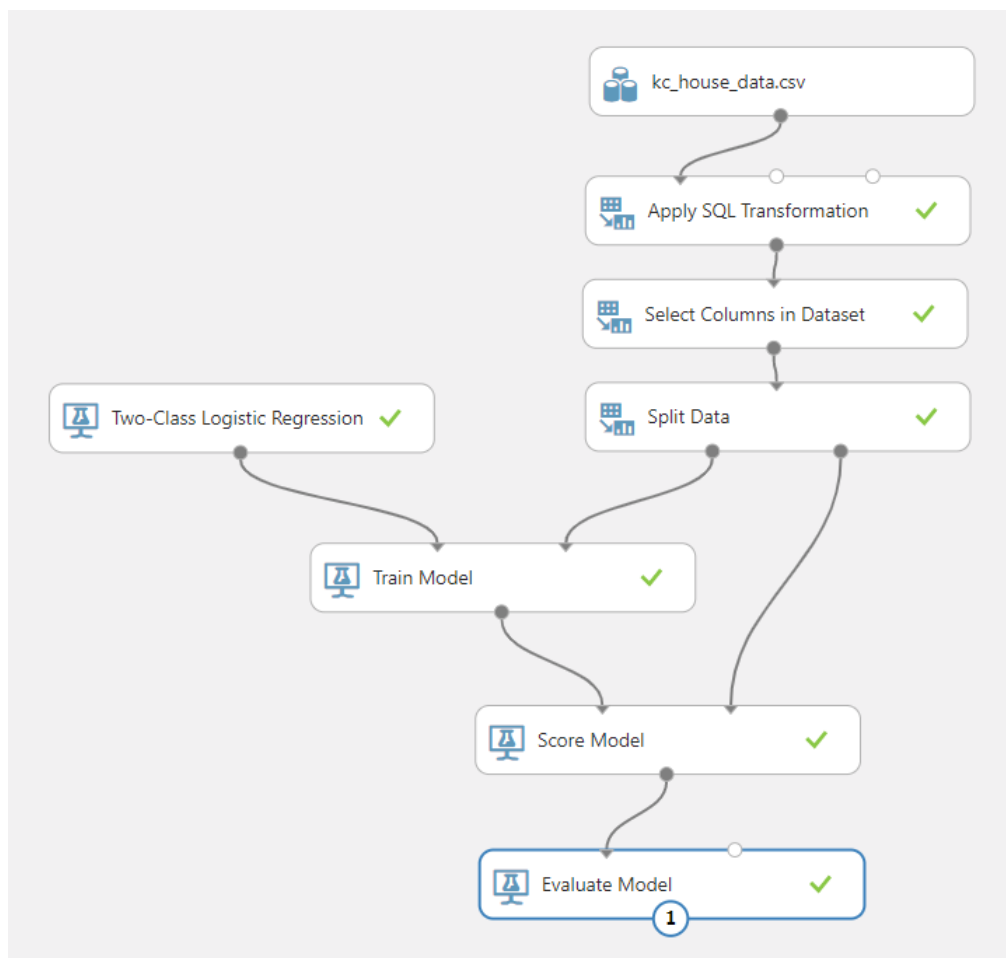
Условие

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

Метод решения

Воспользуемся сначала датасетом проданных домов и попробуем предсказать, является ли данный дом элитным. Элитным будем называть дом, цена которого больше 700000\$, в датасете их около 20%.

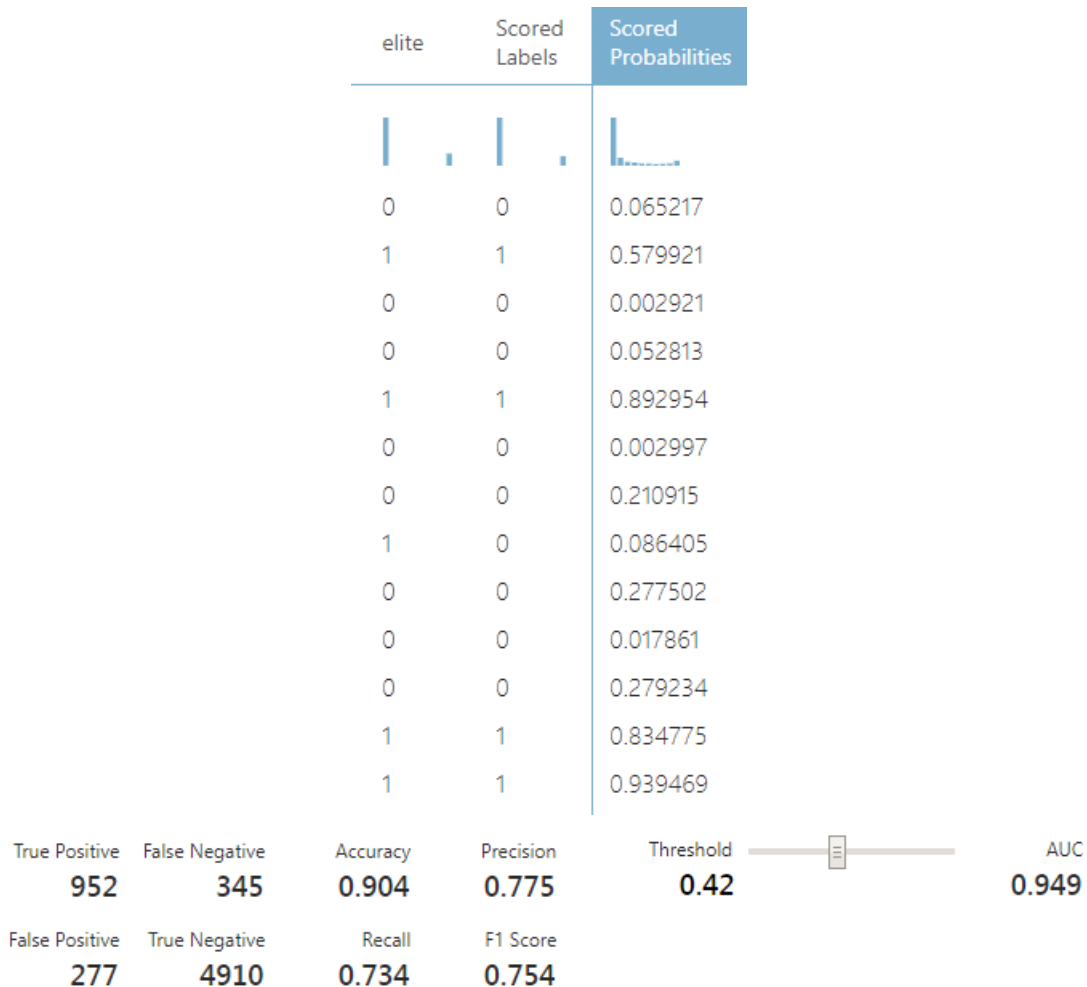
Воспользуемся сначала логистической регрессией.



Сначала мы загружаем датасет, а затем с помощью SQL-запроса добавляет новую колонку *elite*, которую и будем предсказывать. Запрос выглядит следующим образом:

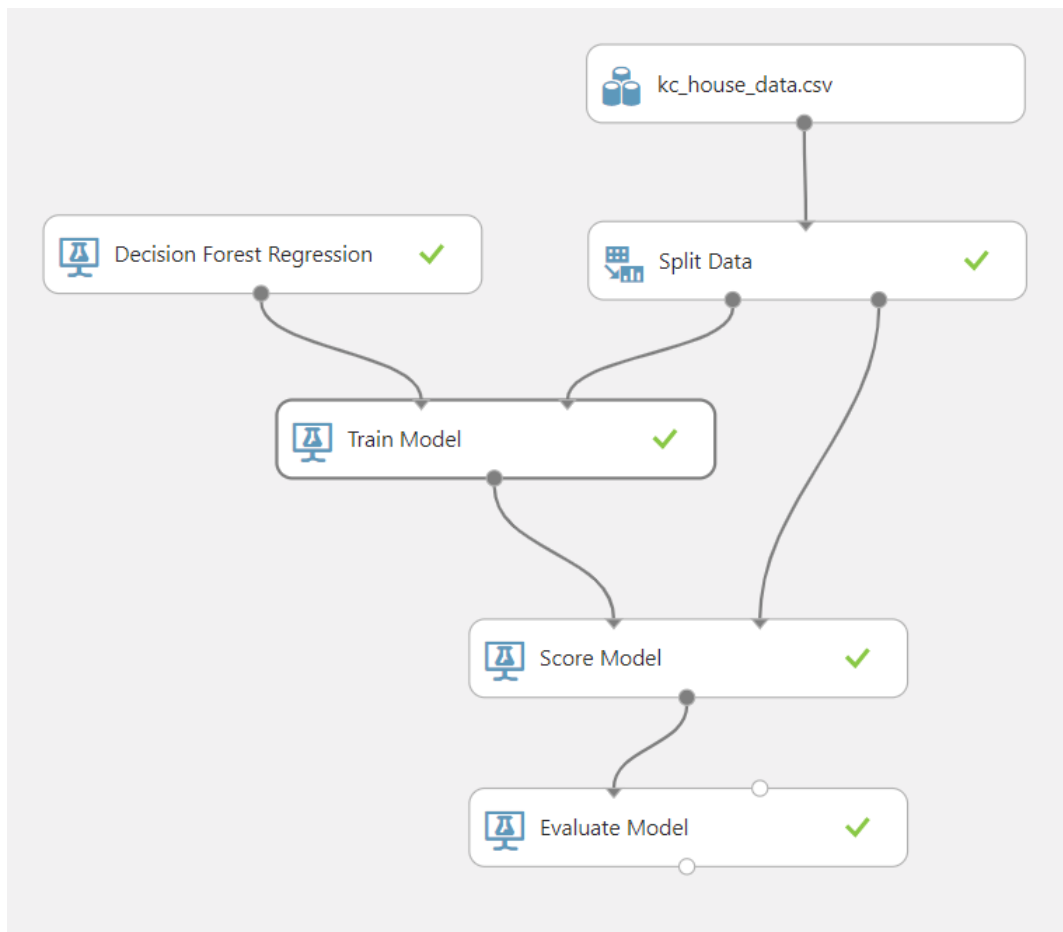
```
select *, price > 700000 as elite from t1;
```

После этого с помощью *Select Columns in Dataset* мы исключаем колонку *price*, так как предсказываемая величина зависит от неё напрямую. Далее разделяем данные на тренировочные и тестирующие в соотношении 70 на 30. После чего наконец обучаем модель и анализируем результаты.






Результаты получились неплохими, так как и precision и recall получились достаточно высокими.

Попробуем теперь предсказать цену на дом с помощью Random Forest Regression.



Также как и в предыдущем примере загружаем датасет, делим его на тренировочную часть и тестирующую, обучаем модель, анализируем её работу.

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
83455.77176	86374.733728	159480.979587	0.369539	0.186994	0.813006





price	Scored Label Mean	Scored Label Standard Deviation
		
335000	333425.070971	42214.043203
378500	323545.833333	58939.931168
660000	556528.94375	158798.418776
220000	380151.5625	111654.309872
625000	689878.522727	184759.754567
442000	452323.958333	114482.59522
215000	240627.571429	18414.570592
700000	648518.223485	87882.9355
124000	221070.795455	62520.558687
205000	221040.208333	53727.072962

Перейдем теперь к рассмотрению корпуса документов (обзоры вин). В данном датасете винам дана оценка от 80 до 100 баллов. Будем считать вино, которые имеет балл выше 90, наилучшим. Таких в выборке около 25%. Попробуем предсказать, является ли вино наилучшим по его обзору.

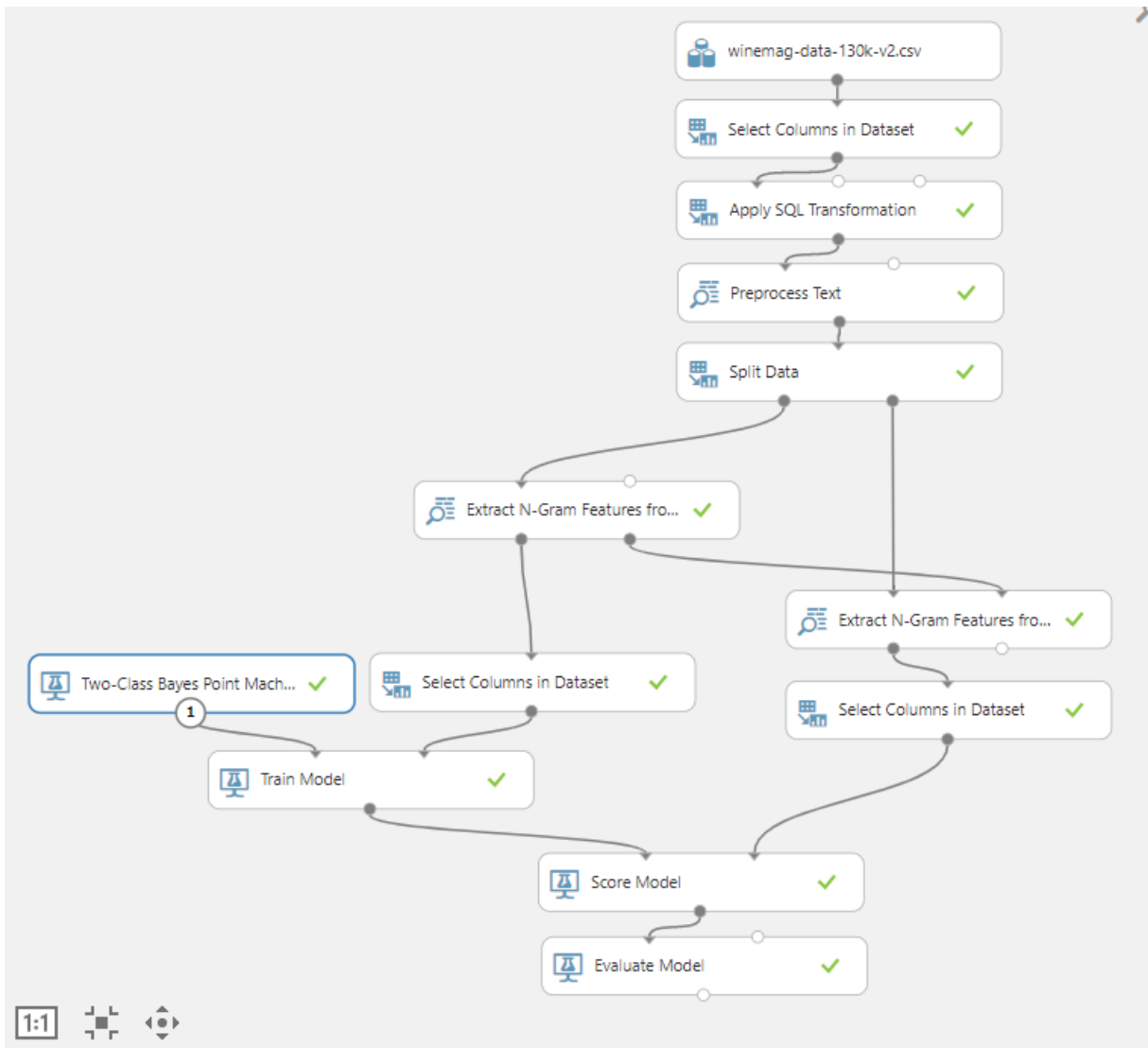
Сначала мы загружаем датасет и убираем все лишние колонки, оставляем только *description* и *points*. После этого с помощью SQL-запроса добавляем колонку *best*:

```
select *, points >= 90 as best from t1;
```

Далее с помощью функции *Process Text* обрабатываем ревью: приводим все слова к нижнему регистру, удаляем знаки препинания и различные окончания, оставляя только корень.

description	points	best	Preprocessed description
			
Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	87	0	aroma include tropical fruit broom brimstone dry herb palate n't overly expressive offer unripen apple citrus dry sage alongside brisk acidity

Потом разбиваем выборку на обучающую и тестирующую, и выделяем из них наиболее значащие фразы с помощью функции *Extract N-Gram Features from Text*. Эта функция добавляет в нашу таблицу дополнительные колонки, соответствующие выделенным N-граммам. С помощью *Select Column in Dataset* оставим только эти новые колонки, они характеризуют частоту появления выделенных N-грамм в обзоре. Будем тренировать нашу модель на них. *Two-Class Bayes Point Machine* показал сравнимую производительность с другими алгоритмами.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
6854	5453	0.665	0.558	0.4	0.703
False Positive	True Negative	Recall	F1 Score		
5436	14750	0.557	0.557		

Как видно, точность оставляет желать лучшего, из вин с оценкой больше 90 алгоритм определил правильно только чуть больше половины. Возможно я выделил слишком мало N-грамм (300) или зря удалил другие признаки.

Выводы

В данной лабораторной работе мы абстрагировались от реализаций алгоритмов и узнали, как происходит разработка решений задач машинного обучения в целом, а также ознакомились с различными метриками на практике.