Московский авиационный институт
(Национальный исследовательский университет)

Факультет: «Информационные технологии и прикладная математика»
Кафедра: 806 «Вычислительная математика и программирование»
Дисциплина: «Машинное обученик»

Лабораторная работа № 1
Azure ML

Студент: Синдюков В.Р.
Группа: М80-308Б
Дата:
Оценка:

Москва, 2019

Постановка задачи

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

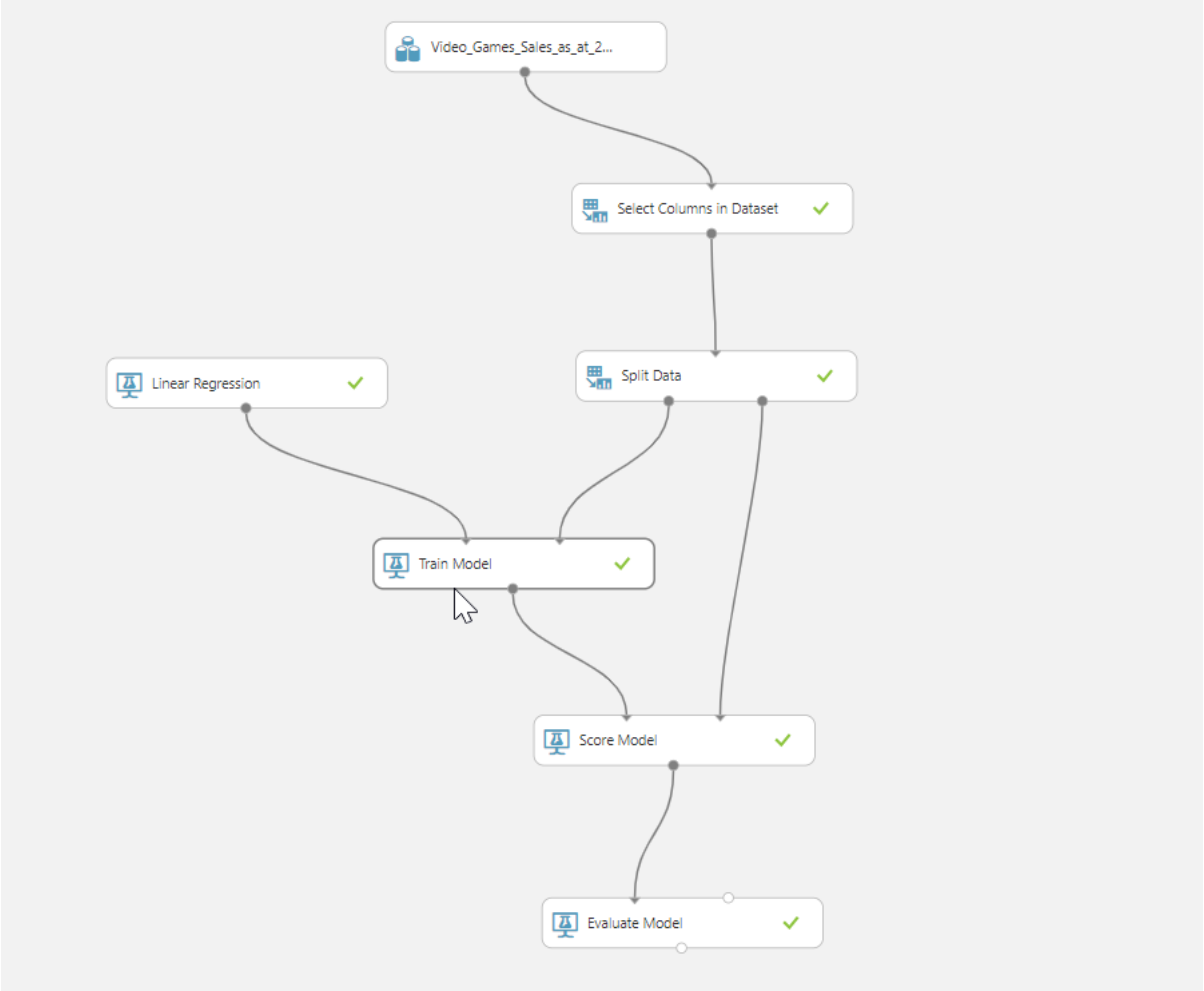
Задача 1

Предсказание мировых продаж видео игры на основании платформы, года выпуска, издателя жанра, американский, европейских и японских продаж.

Мной был использован алгоритм линейной регрессии, так как это один из наиболее простых методов прогнозирования.

Из датасета я выделил 80% данных для обучающей выборки.

Модель эксперимента



Полученные оценки

rows 3344 columns 10

view as

Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Scored Labels
PSV	2013	Adventure	Sony Computer Entertainment Europe	0.14	0.34	0	0.14	0.61	0.620406
PS4	2014	Action	Activision	0.2	0.25	0.02	0.09	0.55	0.560034
PS2	2001	Racing	Electronic Arts	2.02	1.17	0	0.42	3.61	3.609234
XB	2000	Sports	Microsoft Game Studios	0.74	0.21	0	0.04	0.99	0.990038
XB	2002	Sports	Activision	0.15	0.04	0	0.01	0.2	0.200571
PS	1999	Racing	Electronic Arts	0.12	0.08	0	0.01	0.22	0.209811
X360	2009	Sports	Electronic Arts	0.82	0.17	0.01	0.09	1.09	1.089973
Wii	2008	Role-Playing	Sega	0.08	0	0.11	0.01	0.19	0.200074
Wii	2007	Platform	Sega	1.21	1.19	0.04	0.29	2.73	2.730381
NES	1985	Sports	Nintendo	0.18	0.23	1.53	0.02	1.96	1.959347
PS2	2007	Action	Marvelous Interactive	0	0	0.02	0	0.02	0.020343
GBA	2004	Simulation	Ubisoft	0.45	0.17	0	0.01	0.63	0.630504
XOne	2015	Sports	Konami Digital Entertainment	0.04	0.06	0	0.01	0.11	0.110529
GBA	2003	Role-Playing	Atari	0.78	0.29	0	0.02	1.09	1.091291
GC	2005	Shooter	Activision	0.25	0.07	0	0.01	0.33	0.330966
PSP	2011	Role-Playing	Atlus	0.13	0.03	0.11	0.03	0.3	0.299551
GBA	2006	Misc	Konami Digital Entertainment	0.14	0.05	0.07	0.01	0.26	0.271029
PS2	2003	Action	Ubisoft	0.08	0.06	0	0.02	0.16	0.159935

Здесь можно видеть цену, которая была в датасете и предсказанную цену, полученную в ходе расчета.

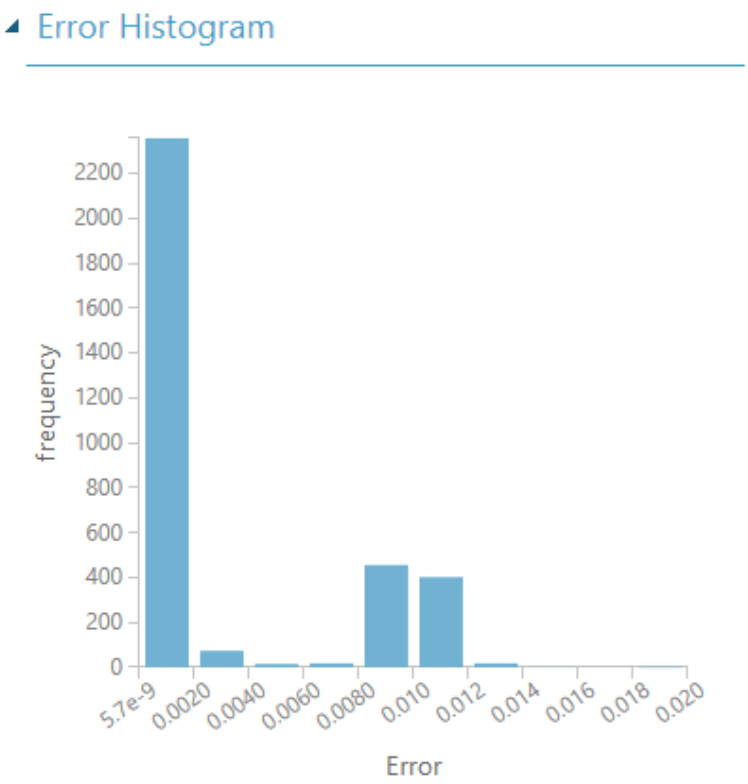
Метрики

Metrics

Mean Absolute Error	0.003118
Root Mean Squared Error	0.005278
Relative Absolute Error	0.005274
Relative Squared Error	0.000013
Coefficient of Determination	0.999987

Средняя абсолютная ошибка, среднеквадратичная ошибка, относительные ошибки и коэффициент смешанной корреляции (детерминированности).

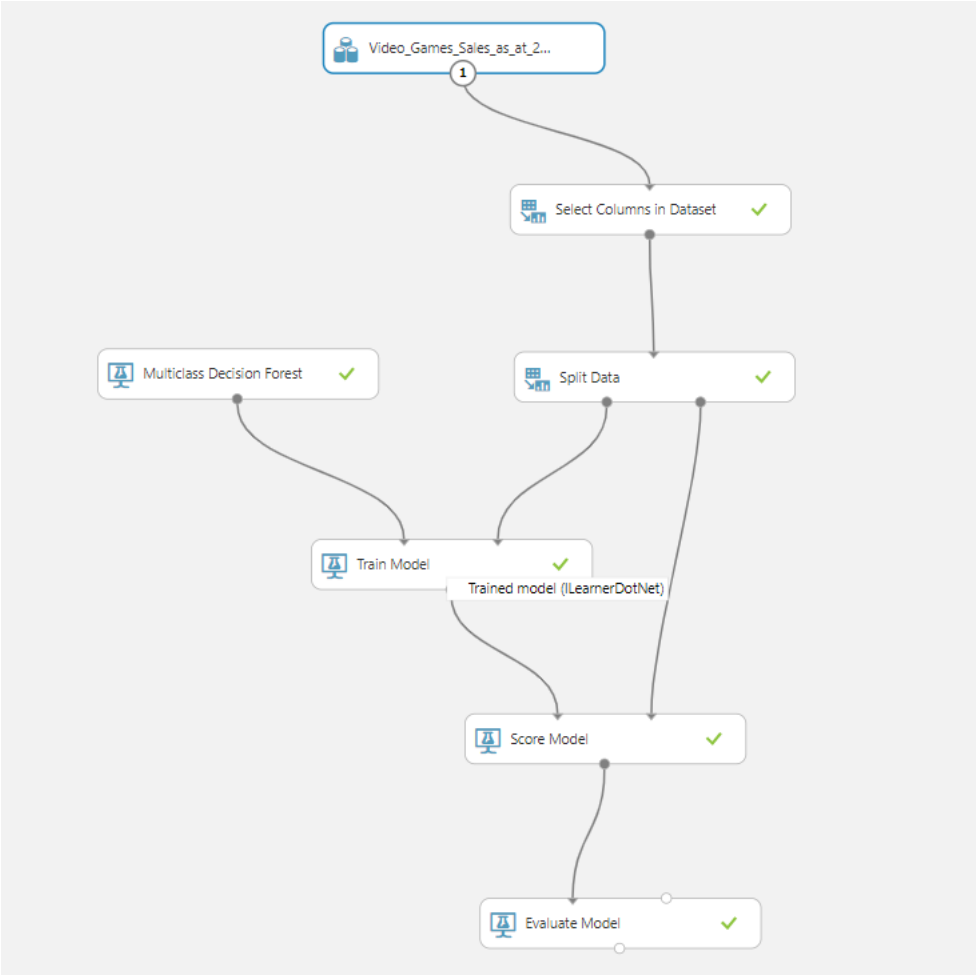
Ошибки и их частоты



Задача 2
Мультиклассовая классификация типа платформы, на которую вышла игра. Из датасета взяты: платформа, год издания, жанр, издатель, продажи, рейтинг.

Для классификации я взял дерево решений, так как оно просто для представления, строится в самом Azure, не требует предварительной подготовки данных и работает со всеми типами.
Выделил 80% датасета для обучения

Модель эксперимента



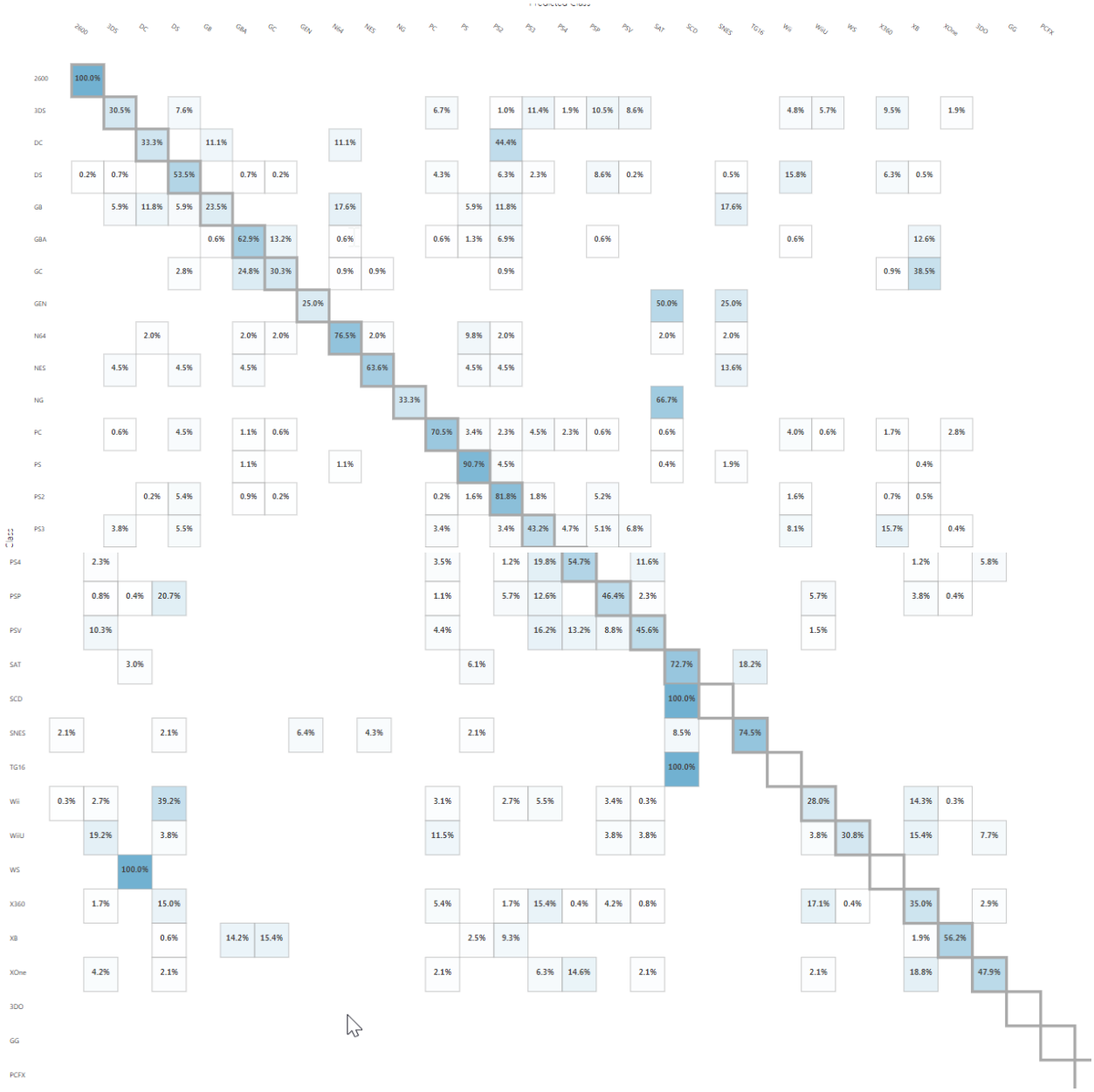
Результат Score Model

rows 3344 columns 40

Scored Probabilities for Class "PSP"	Scored Probabilities for Class "PSV"	Scored Probabilities for Class "SAT"	Scored Probabilities for Class "SCD"	Scored Probabilities for Class "SNES"	Scored Probabilities for Class "Wii"	Scored Probabilities for Class "WiiU"	Scored Probabilities for Class "WS"	Scored Probabilities for Class "X360"	Scored Probabilities for Class "XB"	Scored Probabilities for Class "XOne"	Scored Labels
0	0.241542	0	0	0	0.125	0.010542	0	0.128614	0	0.04006	PSV
0	0	0	0	0	0	0	0	0.028125	0	0	PS3
0	0	0	0	0	0	0	0	0	0	0	PS2
0.008929	0	0	0	0	0.017857	0	0	0.035714	0.026786	0	PS
0	0	0	0	0	0	0	0	0	0.627161	0	XB
0	0	0	0	0	0	0	0	0	0	0	PS
0.003472	0	0	0	0	0.098611	0	0	0.353009	0	0	X360
0.098958	0.020833	0	0	0	0.1875	0	0	0	0	0	DS
0.007353	0	0	0	0	0.358456	0	0	0.225	0.020833	0	Wii
0	0	0.125	0	0.022727	0.005682	0.005682	0	0	0	0	NES
0.014706	0	0	0	0	0.155282	0	0	0	0	0	DS
0	0	0	0	0	0	0	0	0	0.027778	0	GBA

В результате мы имеем оцененные вероятности каждого класса для всех позиций и класс, который наиболее вероятен.

Матрица ошибок



Метрики

Metrics

Overall accuracy	0.559211
Average accuracy	0.971562
Micro-averaged precision	0.559211
Macro-averaged precision	NaN
Micro-averaged recall	0.559211
Macro-averaged recall	NaN

Общая точность, средняя точность, микро- и макро- усредненная точность и полнота.

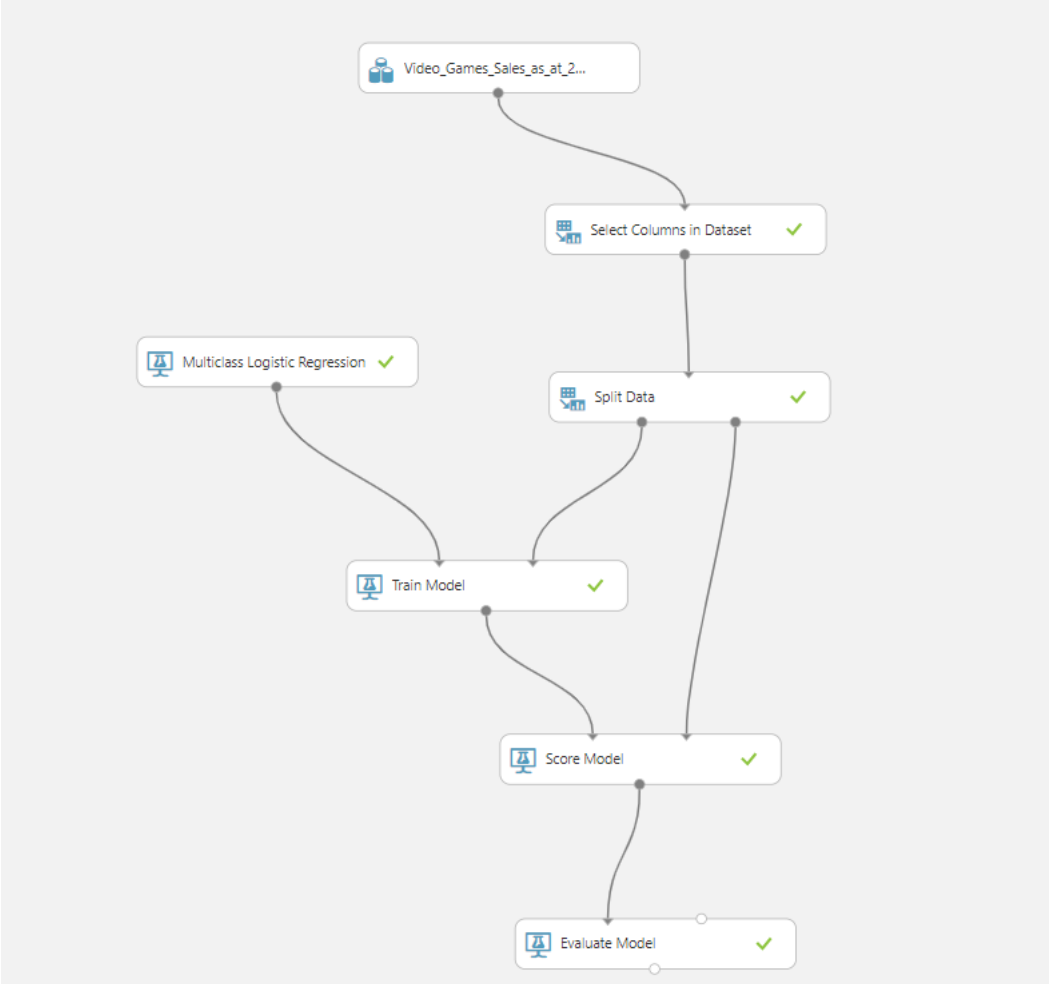
Задача 3
Классификация игр по рейтингу. Из датасета взял платформу, дату выхода, жанр, издателя, продажи, рейтинг.

Выделил 80% на обучение.

Для классификации выбрал мультиклассовую логическую регрессию, так как она отлично подходит для решеная подобных задач:
- Какой тип крови у человека, учитывая результаты различных диагностических тестов?
- В какой стране фирма будет располагать офисом, учитывая характеристики фирмы и различных стран-кандидатов?

В данном случае я хочу понять рейтинг игры в зависимости от издателя, года выпуска, жара и продаж.

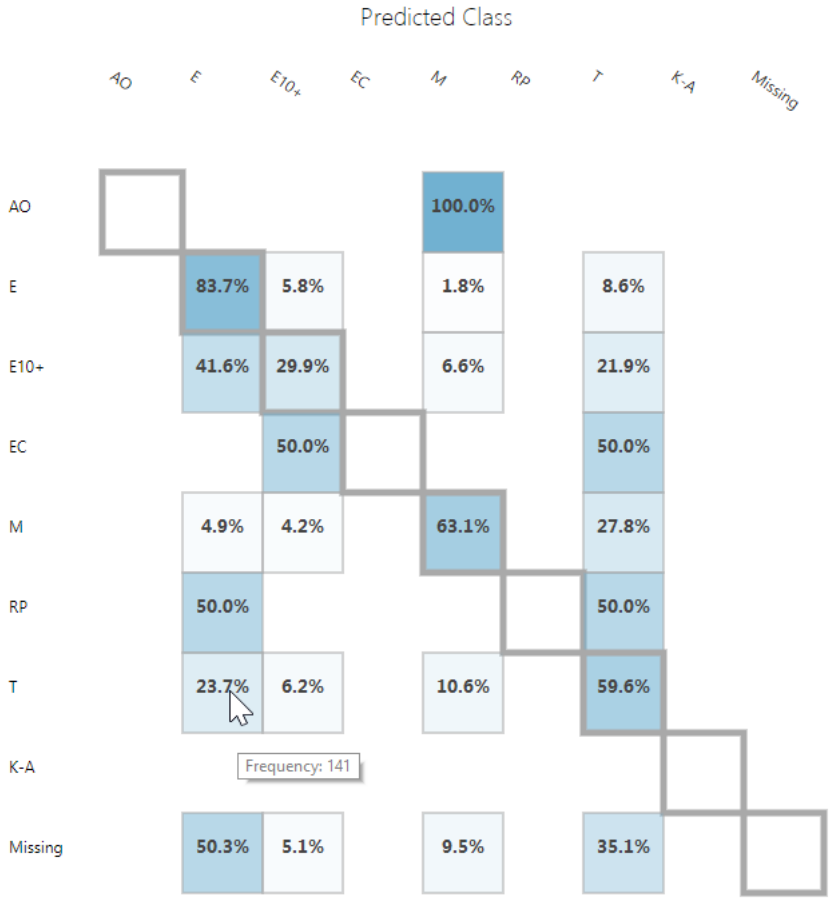
Модель эксперимента



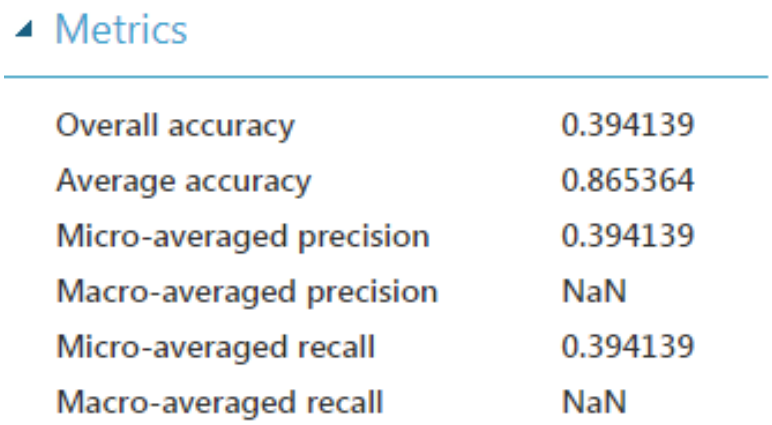
Результат обучения

Rating	Scored Probabilities for Class "E"	Scored Probabilities for Class "E10+"	Scored Probabilities for Class "EC"	Scored Probabilities for Class "K-A"	Scored Probabilities for Class "M"	Scored Probabilities for Class "RP"	Scored Probabilities for Class "T"	Scored Labels
E	0.147687	0.155226	0.000893	0.000649	0.428155	0.000245	0.267145	M
	0.037631	0.282124	0.000535	0.000389	0.319798	0.000147	0.359378	T
T	0.833279	0.012271	0.000219	0.000159	0.006145	0.00006	0.147867	E
E	0.893394	0.005623	0.000142	0.000103	0.008374	0.000039	0.092326	E
T	0.664295	0.00444	0.000157	0.000114	0.004364	0.000043	0.326588	E
	0.909581	0.017284	0.000205	0.000149	0.004386	0.000056	0.068339	E
T	0.777547	0.104178	0.000292	0.000212	0.010877	0.00008	0.106814	E
T	0.102576	0.231068	0.000886	0.00059	0.062114	0.000223	0.602543	T
E	0.557698	0.367361	0.000371	0.000247	0.002915	0.000093	0.071315	E
	0.914003	0.02453	0.000142	0.000103	0.001314	0.000039	0.059869	E
	0.116412	0.296475	0.000784	0.00057	0.219882	0.000215	0.365662	T

Матрица ошибок



Можно видеть, что классификация оказалась не совсем точной. Это потом что, к сожалению, в некоторых столбцах присутствовали пустые строки(их было около половины)



Выводы

Эта лабораторная работы была полезна, так как благодаря ей можно довольно быстро и просто ознакомится с основными алгоритмами машинного обучения. Также для написания самих алгоритмов хорошо бы посмотреть как они работают на примере готовых данных и какие результаты выдают, чтобы можно было уже сверять со своими и вносить необходимые правки.