

# **Отчет по лабораторной работе №1 по курсу «Искусственный интеллект»**

Выполнил студент группы М8О-3086-16 Никитин Андрей

## **Тема:**

Работа с Azure Machine Learning Studio

## **Задача:**

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

## **Оборудование студента:**

Ноутбук Lenovo ThinkPad 13, процессор Intel® Core™ i5-7200U CPU 1.70 GHz, память 8ГБ, 64-разрядная система.

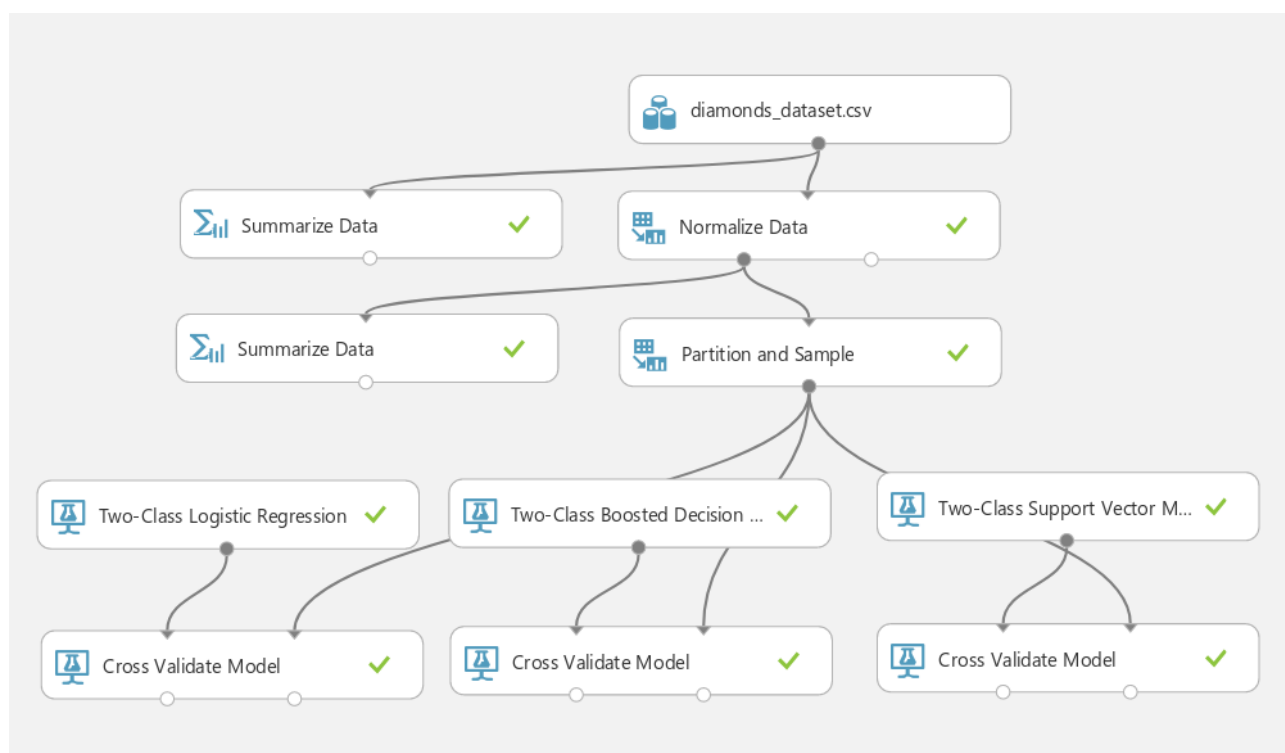
## **Программное обеспечение:**

ОС Linux Mint 19, Mozilla Firefox 66.0.2

## **Ход работы:**

В данной лабораторной работе используется датасет бриллиантов, описанный в предыдущей лабораторной работе, для которого поставлена задача классификации на два класса: бриллианты дороже 3000\$ и бриллианты стоимостью до 3000\$.

Работа с данными на платформе Azure ML осуществляется посредством экспериментов. Эксперименты открывают широкий доступ к алгоритмам, которые используются в машинном обучении, начиная с загрузки данных и сбора статистики, заканчивая тестированием уже обученной модели. Все это реализовано в виде интерактивной схемы:













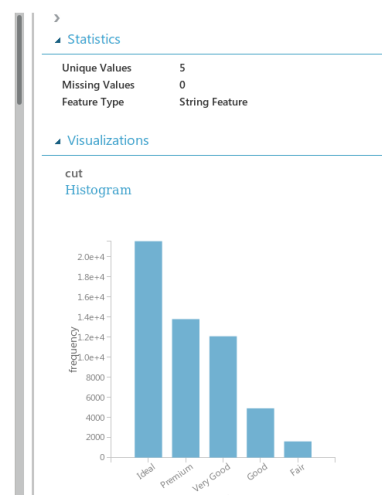
Разберем каждую операцию в данной схеме. В первую очередь выполняется загрузка данных. Данная платформа позволяет сразу же визуализировать загруженный датасет:

Diamonds Price Prediction > diamonds\_dataset.csv > dataset

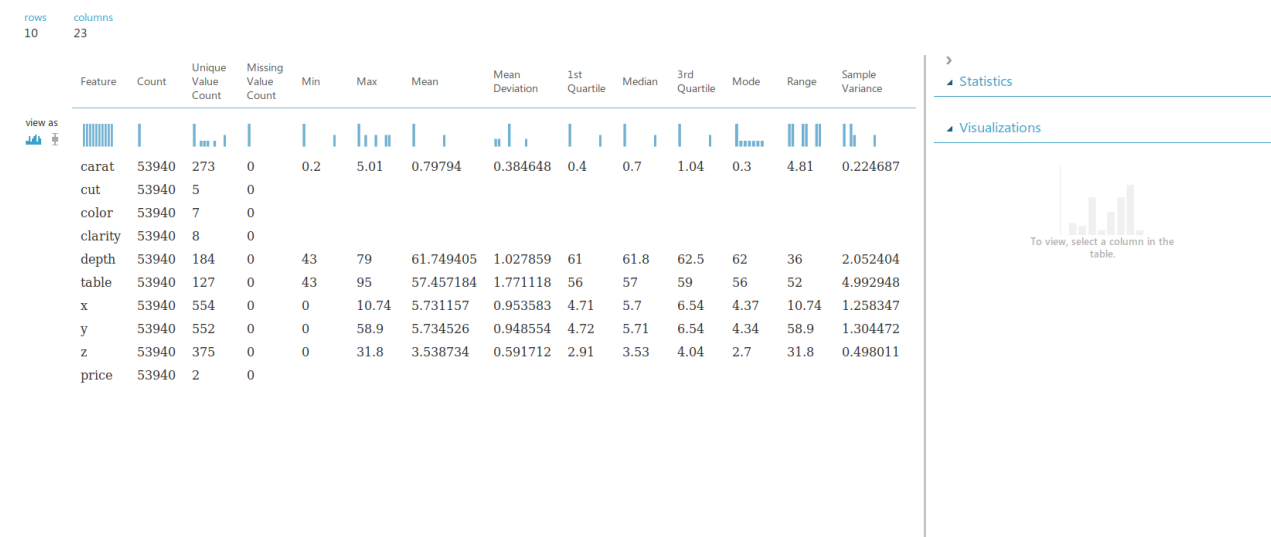
rows  
53940

columns  
10

	carat	cut	color	clarity	depth	table	x	y	z	price
view as										
0.23	Ideal	E	SI2	61.5	55	3.95	3.98	2.43	<= 3K	
0.21	Premium	E	SI1	59.8	61	3.89	3.84	2.31	<= 3K	
0.23	Good	E	VS1	56.9	65	4.05	4.07	2.31	<= 3K	
0.29	Premium	I	VS2	62.4	58	4.2	4.23	2.63	<= 3K	
0.31	Good	J	SI2	63.3	58	4.34	4.35	2.75	<= 3K	
0.24	Very Good	J	VVS2	62.8	57	3.94	3.96	2.48	<= 3K	
0.24	Very Good	I	VVS1	62.3	57	3.95	3.98	2.47	<= 3K	
0.26	Very Good	H	SI1	61.9	55	4.07	4.11	2.53	<= 3K	
0.22	Fair	E	VS2	65.1	61	3.87	3.78	2.49	<= 3K	
0.23	Very Good	H	VS1	59.4	61	4	4.05	2.39	<= 3K	
0.3	Good	J	SI1	64	55	4.25	4.28	2.73	<= 3K	
0.23	Ideal	J	VS1	62.8	56	3.93	3.9	2.46	<= 3K	
0.22	Premium	F	SI1	60.4	61	3.88	3.84	2.33	<= 3K	
0.31	Ideal	J	SI2	62.2	54	4.35	4.37	2.71	<= 3K	
0.2	Premium	E	SI2	60.2	62	3.79	3.75	2.27	<= 3K	
0.32	Premium	E	I1	60.9	58	4.38	4.42	2.68	<= 3K	
0.3	Ideal	I	SI2	62	54	4.31	4.34	2.68	<= 3K	
0.3	Good	J	SI1	63.4	54	4.23	4.29	2.7	<= 3K	



При решении задач машинного обучения бывает часто полезно посмотреть на статистические характеристики загруженного датасета. Для этого доступна операция «Summarize Data», которая вычисляет математическое ожидание, стандартное отклонение, минимум, максимум и другие полезные величины:



Стоит заметить, что поле «Missing values count» для всех признаков равно 0, поэтому этап обработки пропущенных значений можно пропустить. Однако, так как некоторые признаки чувствительны к масштабу, неплохо бы сделать нормализацию признаков (Операция «Normalize Data»). В данном конкретном случае нормализация выполняется посредством вычета среднего и деления на стандартное отклонение. Результат нормализации можно увидеть ниже:



У всех количественных признаков теперь среднее ноль, а значит нормализация прошла успешно. Для тестирования алгоритмов будет использоваться кросс-валидация по пяти блокам. Для ее работы необходимо создать разбиение нашего датасета с помощью операции «Partition and Sample». Эта операция делит датасет на пять равных частей, причем случайным образом. Как уже говорилось для тестирования модели будет использоваться кросс-валидация. Её суть заключается в том, что исходный датасет разбивается на N частей (Обычно берут 5 или 10), затем каждый блок поочередно используется для тестирования,

а остальные для обучения. Хотя при таком подходе выполняется N обучений, данная техника имеет множество достоинств. Например, модель тестируется на большем количестве данных. Также стоит отметить, что кросс-валидация позволяет заметить склонность полученной модели к переобучению, что тоже является очень важным параметром при оценке.

Теперь, когда все приготовления завершены, можно приступить к непосредственному обучению модели. Для данной лабораторной работы, были использованы следующие алгоритмы классификации: Two-Class Support Vector Machine, Two-Class Logistic Regression и Two-Class Boosted Decision Tree. Данный выбор обусловлен тем, что данные алгоритмы мне знакомы и они разбирались на лекциях (За исключением последнего). Рассмотрим поочередно результаты их работы.

Метод опорных векторов (Two-Class Support Vector Machine) показал неплохие но не самые лучшие результаты. Стоит отметить, что понижение коэффициента для L1-регуляризации лучше сказалось на точности полученной модели (При lambda равной 0.1 точность была примерно 0.75, когда для lambda 0.0001 точность достигает 0.97). На первом скриншоте видно к какому блоку была отнесена каждая запись и какой класс был предсказан для него. Оценка алгоритма же представлена на втором.

Diamonds Price Prediction > Cross Validate Model > Scored results

rows  
53940

columns  
13

	Fold Assignments	carat	cut	color	clarity	depth	table	x	y	z	price	Scored Labels
view as												
3	-1.198168	Ideal	E	SI2	-0.174092	-1.099672	-1.587837	-1.536196	-1.571129	<= 3K	<= 3K	
3	-1.240361	Premium	E	SI1	-1.360738	1.585529	-1.641325	-1.658774	-1.741175	<= 3K	<= 3K	
2	-1.198168	Good	E	VS1	-3.385019	3.375663	-1.498691	-1.457395	-1.741175	<= 3K	<= 3K	
3	-1.071587	Premium	I	VS2	0.454133	0.242928	-1.364971	-1.317305	-1.28772	<= 3K	<= 3K	
2	-1.029394	Good	J	SI2	1.082358	0.242928	-1.240167	-1.212238	-1.117674	<= 3K	<= 3K	
3	-1.177071	Very Good	J	VVS2	0.733344	-0.204605	-1.596752	-1.553707	-1.500277	<= 3K	<= 3K	
1	-1.177071	Very Good	I	VVS1	0.384331	-0.204605	-1.587837	-1.536196	-1.514447	<= 3K	<= 3K	
0	-1.134878	Very Good	H	SI1	0.10512	-1.099672	-1.480862	-1.422373	-1.429424	<= 3K	<= 3K	
4	-1.219265	Fair	E	VS2	2.338808	1.585529	-1.659155	-1.711308	-1.486106	<= 3K	<= 3K	
1	-1.198168	Very Good	H	VS1	-1.63995	1.585529	-1.543264	-1.474906	-1.627811	<= 3K	<= 3K	
2	-1.050491	Good	J	SI1	1.570978	-1.099672	-1.320398	-1.273527	-1.146015	<= 3K	<= 3K	
4	-1.198168	Ideal	I	VS1	0.733344	-0.652139	-1.605667	-1.60624	-1.528618	<= 3K	<= 3K	

Statistics

Visualizations

To view, select a column in the table.

rows7

columns10

view as

Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
0	10788	SVM (Pegasos-Linear)	0.97145	0.967357	0.966944	0.96715	0.996752	0.072783	89.368248
1	10788	SVM (Pegasos-Linear)	0.975899	0.974392	0.970694	0.972539	0.996773	0.071424	89.586058
2	10788	SVM (Pegasos-Linear)	0.972191	0.97117	0.96453	0.967839	0.997063	0.068407	90.004316
3	10788	SVM (Pegasos-Linear)	0.974231	0.973124	0.967301	0.970204	0.997473	0.063836	90.671776
4	10788	SVM (Pegasos-Linear)	0.974045	0.970717	0.97112	0.970918	0.997571	0.064917	90.555188
Mean	53940	SVM (Pegasos-Linear)	0.973563	0.971352	0.968118	0.96973	0.997126	0.068273	90.037117
Standard Deviation	53940	SVM (Pegasos-Linear)	0.001767	0.002682	0.002764	0.002223	0.000383	0.003912	0.575132

>

Statistics

Visualizations

To view, select a column in the table.

Интересно посмотреть на параметры accuracy, precision и recall. Данные значения являются базовыми метриками качества. В задачах бинарной классификации часто используют так называемую ROC-кривую для оценки модели. В данном случае операция «Cross Validation Model» её не строит, однако вычисляет другую интересную величину — AUC (area under curve). Чем ближе это значение к 1 тем лучше работает классификатор, значения близкие к 0.5 означают, что наш классификатор работает по принципу подбрасывания монетки.

Выполним те же самые действия для логистической регрессии (Коэффициенты регуляризации взяты по 1, точности оптимизации  $10^{-7}$ ):

rows

columns

53940

13

view as

Fold Assignments

carat

cut

color

clarity

depth

table

x

y

z

price

Scored Labels

3

-1.198168

Ideal

E

SI2

-0.174092

-1.099672

-1.587837

-1.536196

-1.571129

<= 3K

<= 3K

3

-1.240361

Premium

E

SI1

-1.360738

1.585529

-1.641325

-1.658774

-1.741175

<= 3K

<= 3K

2

-1.198168

Good

E

VS1

-3.385019

3.375663

-1.498691

-1.457395

-1.741175

<= 3K

<= 3K

3

-1.071587

Premium

I

VS2

0.454133

0.242928

-1.364971

-1.317305

-1.28772

<= 3K

<= 3K

2

-1.029394

Good

J

SI2

1.082358

0.242928

-1.240167

-1.212238

-1.117674

<= 3K

<= 3K

3

-1.177071

Very Good

J

VVS2

0.733344

-0.204605

-1.596752

-1.553707

-1.500277

<= 3K

<= 3K

1

-1.177071

Very Good

I

VVS1

0.384331

-0.204605

-1.587837

-1.536196

-1.514447

<= 3K

<= 3K

0

-1.134878

Very Good

H

SI1

0.10512

-1.099672

-1.480862

-1.422373

-1.429424

<= 3K

<= 3K

4

-1.219265

Fair

E

VS2

2.338808

1.585529

-1.659155

-1.711308

-1.486106

<= 3K

<= 3K

1

-1.198168

Very Good

H

VS1

-1.63995

1.585529

-1.543264

-1.474906

-1.627811

<= 3K

<= 3K

2

-1.050491

Good

J

SI1

1.570978

-1.099672

-1.320398

-1.273527

-1.146015

<= 3K

<= 3K

4

-1.198168

Ideal

I

VS1

0.733344

-0.652139

-1.605667

-1.60624

-1.528618

<= 3K

<= 3K



>

Statistics

Visualizations

To view, select a column in the table.

rows  
7columns  
10

	Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
view as  										
	0	10788	Logistic Regression	0.974972	0.973633	0.96865	0.971135	0.997328	0.087478	87.22171
	1	10788	Logistic Regression	0.978495	0.97918	0.971748	0.97545	0.997165	0.085069	87.596455
	2	10788	Logistic Regression	0.975806	0.976493	0.967521	0.971987	0.997587	0.084557	87.644362
	3	10788	Logistic Regression	0.976919	0.979022	0.967514	0.973234	0.997866	0.083785	87.756745
	4	10788	Logistic Regression	0.977197	0.976623	0.972159	0.974386	0.997941	0.083814	87.805895
	Mean	53940	Logistic Regression	0.976678	0.97699	0.969518	0.973238	0.997577	0.084941	87.605033
	Standard Deviation	53940	Logistic Regression	0.001351	0.002268	0.002275	0.001746	0.000335	0.001517	0.230173

&gt;

Statistics

Visualizations



И для бустинга над решающими деревьями (100 деревьев, максимум 20 листьев на дерево):

rows  
53940columns  
13



	Fold Assignments	carat	cut	color	clarity	depth	table	x	y	z	price	Scored Labels
view as  												
	3	-1.198168	Ideal	E	SI2	-0.174092	-1.099672	-1.587837	-1.536196	-1.571129	<= 3K	<= 3K
	3	-1.240361	Premium	E	SI1	-1.360738	1.585529	-1.641325	-1.658774	-1.741175	<= 3K	<= 3K
	2	-1.198168	Good	E	VS1	-3.385019	3.375663	-1.498691	-1.457395	-1.741175	<= 3K	<= 3K
	3	-1.071587	Premium	I	VS2	0.454133	0.242928	-1.364971	-1.317305	-1.28772	<= 3K	<= 3K
	2	-1.029394	Good	J	SI2	1.082358	0.242928	-1.240167	-1.212238	-1.117674	<= 3K	<= 3K
	3	-1.177071	Very Good	J	VVS2	0.733344	-0.204605	-1.596752	-1.553707	-1.500277	<= 3K	<= 3K
	1	-1.177071	Very Good	I	VVS1	0.384331	-0.204605	-1.587837	-1.536196	-1.514447	<= 3K	<= 3K
	0	-1.134878	Very Good	H	SI1	0.10512	-1.099672	-1.480862	-1.422373	-1.429424	<= 3K	<= 3K
	4	-1.219265	Fair	E	VS2	2.338808	1.585529	-1.659155	-1.711308	-1.486106	<= 3K	<= 3K
	1	-1.198168	Very Good	H	VS1	-1.63995	1.585529	-1.543264	-1.474906	-1.627811	<= 3K	<= 3K
	2	-1.050491	Good	J	SI1	1.570978	-1.099672	-1.320398	-1.273527	-1.146015	<= 3K	<= 3K
	4	-1.198168	Ideal	I	VS1	0.733344	-0.652139	-1.605667	-1.60624	-1.528618	<= 3K	<= 3K

&gt;

Statistics

Visualizations

rows  
7columns  
10

	Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
view as  										
	0	10788	FastTree (Boosted Trees) Classification	0.980349	0.977598	0.977181	0.977389	0.998327	0.0549	91.980542
	1	10788	FastTree (Boosted Trees) Classification	0.981275	0.978101	0.979338	0.978719	0.998552	0.052225	92.385332
	2	10788	FastTree (Boosted Trees) Classification	0.98211	0.98051	0.978205	0.979356	0.998376	0.052057	92.393318
	3	10788	FastTree (Boosted Trees) Classification	0.982017	0.978655	0.97991	0.979282	0.998043	0.053363	92.202268
	4	10788	FastTree (Boosted Trees) Classification	0.981275	0.97723	0.980885	0.979054	0.998479	0.050997	92.580486
	Mean	53940	FastTree (Boosted Trees) Classification	0.981405	0.978419	0.979104	0.97876	0.998355	0.052708	92.308389
	Standard Deviation	53940	FastTree (Boosted Trees) Classification	0.000711	0.001286	0.001448	0.000806	0.000196	0.001485	0.226889

&gt;

Statistics

Visualizations



Хотя последний алгоритм показал лучшие результаты, я бы отдал предпочтение линейным методам классификации, которые показали результат немного хуже, но за счет своей простоты их легче реализовать, да и обучаются они быстрее, чем большое количество деревьев деревьев.

### **Вывод:**

Таким образом, я познакомился с платформой для решения задач машинного обучения Azure Machine Learning Studio, которая действительно предоставляет широкие возможности для анализа и обработки данных, построения предсказательных моделей и их тестирования. Все это сопровождается интуитивно понятным графическим интерфейсом, что делает работу более приятной.