3 Azure ML Studio

3.1 Модель логистической регрессии

Для обучения модели логистической регрессии используется блок `Two-Class Logistic Regression` из раздела `Machine Learning`. Единственный параметр, который нас будет интересовать – `Random number seed`. Остальные параметры остаются заданными по умолчанию.

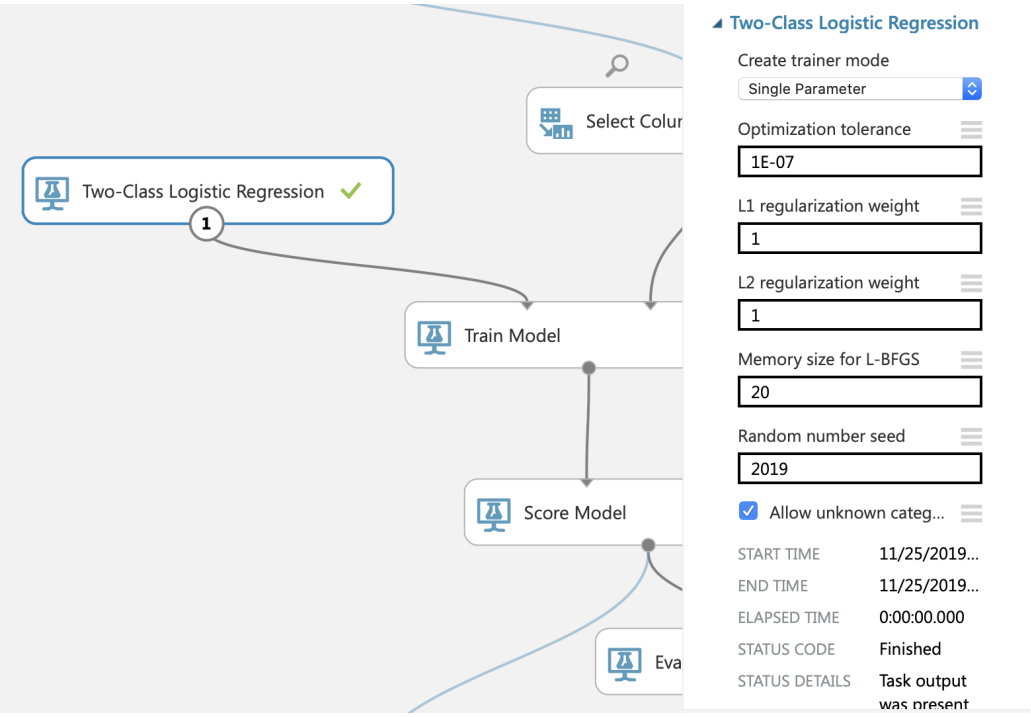

Рис. 1: Блок логистической регрессии.

Блок `Train Model` все также отвечает за обучение модели. На вход подаются данные и выбранный метод машинного обучения. В параметрах данного блока необходимо выбрать столбец данных, соответствующий отклику. Для логистической регрессии это должны быть два класса 0 и 1 или -1 и 1.

После запуска модели, полученные значения коэффициентов уравнения линейной регрессии можно посмотреть в параметрах блока `Train Model`, пункт `Visualize`. Параметр `Bias` (смещение) соответствует коэффициенту θ_0, а названия столбцов данных соответствующим коэффициентам $\theta_1, \ldots, \theta_p$.

Feature Weights

Feature	Weight
chocolate	1.609
hard	-0.686988
Bias	-0.640604
peanutyalmondy	0.313946
caramel	0.222867
bar	0.149035
sugarpercent	0.139294
crispedricewafer	0.0292136
fruity	0
nougat	0
pluribus	0
pricepercent	0

Рис. 2: Параметры модели логистической регрессии.

3.2 Задача классификации

Для решения задачи классификации необходимы данные. В качестве данных могут выступать либо данные в формате CSV, либо это могут быть введенные вручную значения с помощью блока `Enter Data Manually`. После запуска эксперимента `Run` результаты классификации доступны в пункте `Visualize` блока `Score Model`.

К набору данных добавляются колонки `Scored Probabilities` и `Scored Labels`. В первой указана вероятность отнесения объекта к положительному классу, а во второй – результат бинарной классификации. Положительный класс назначается, если вероятность больше или равна 0.5.

bar	pluribus	sugarpercent	pricepercent	Y	Scored Labels	Scored Probabilities
0	1	0.647364	0.767	0	0	0.266198
0	0	0.418	0.325	0	1	0.584054
0	0	0.162	0.116	0	0	0.349992
0	1	0.604	0.755	1	0	0.36446
0	0	0.87656	0.5654	1	1	0.748721
0	0	0.313	0.511	0	1	0.733305
0	1	0.174	0.011	0	1	0.729411
1	0	0.465	0.325	1	1	0.765336
0	1	0.313	0.255	0	0	0.354905

Рис. 3: Результаты классификации.

4 ROC-анализ

За оценку модели отвечает блок `Evaluate Model`, подключаемый к `Score Model`, при этом, к блоку `Score Model` должны быть подключены тестовые данные, содержащие все предикторы и отклик. После обучения модели и запуска доступны:

- `Confusion matrix` (матрица ошибок):

Матрица ошибок		Верный класс	
		+	−
Прогноз	+	TP	FP
	−	FN	TN

- `Precision` (точность) – это доля объектов, действительно являющихся положительными к тем, что названы положительными в результате классификации:
$$\texttt{Precision} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FP}}.$$
- `Recall` (полнота) – это доля объектов, классифицированных, как положительные, к тем, что действительно являются положительными. Также называется долей истинно положительных примеров `TPR` (True Positives Rate):
$$\texttt{Recall} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FN}}.$$

- `AUC` (площадь под кривой).

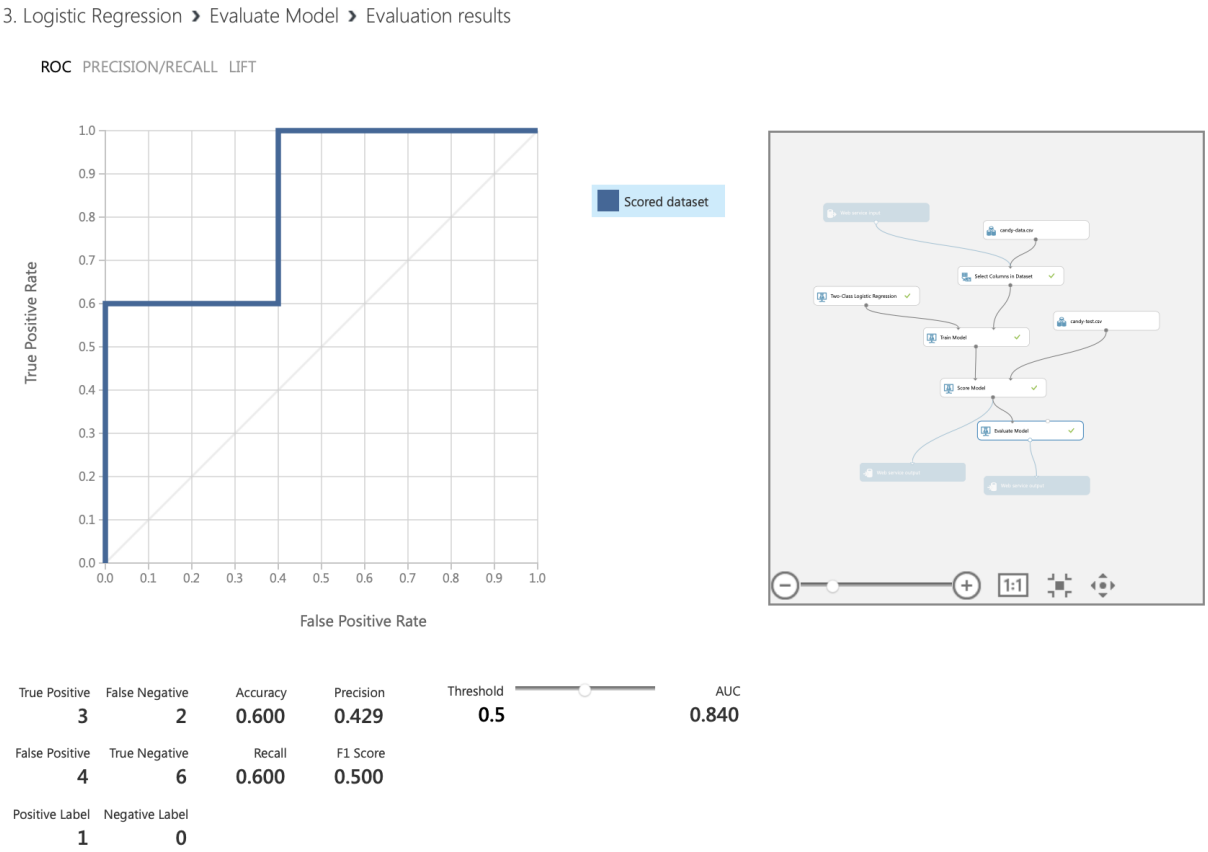

Рис. 4: Оценка модели.

Ползунок для значения `Threshold` позволяет изменять порог отсечения, тем самым влияя на результат бинарной классификации.

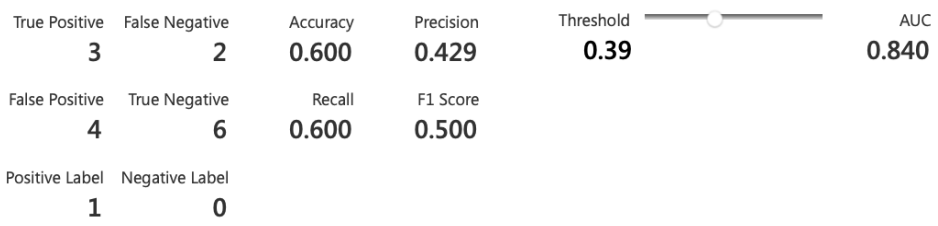

Рис. 5: Изменение порога отсечения.