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: К. В. Лукашкин
Группа: М8О-308Б

Москва, 2019

Постановка задачи

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

Решение

В данной лабораторной работе я работал с датасетом из нулевой лабораторной – история акций компании NASDAQ Composite (^IXIC) с 1999 года https://finance.yahoo.com/quote/%5EIXIC/history

Было использовано 4 алгоритма – Decision Forest Regresssion, Logistic Regression, Two-class Decision Forest Classification, Two-class Logistic Regression.

Для того чтобы к датасету можно было применить алгоритмы классификации, был также введен дополнительный столбец profit, который отображает была ли получена прибыль в данный день торгов.

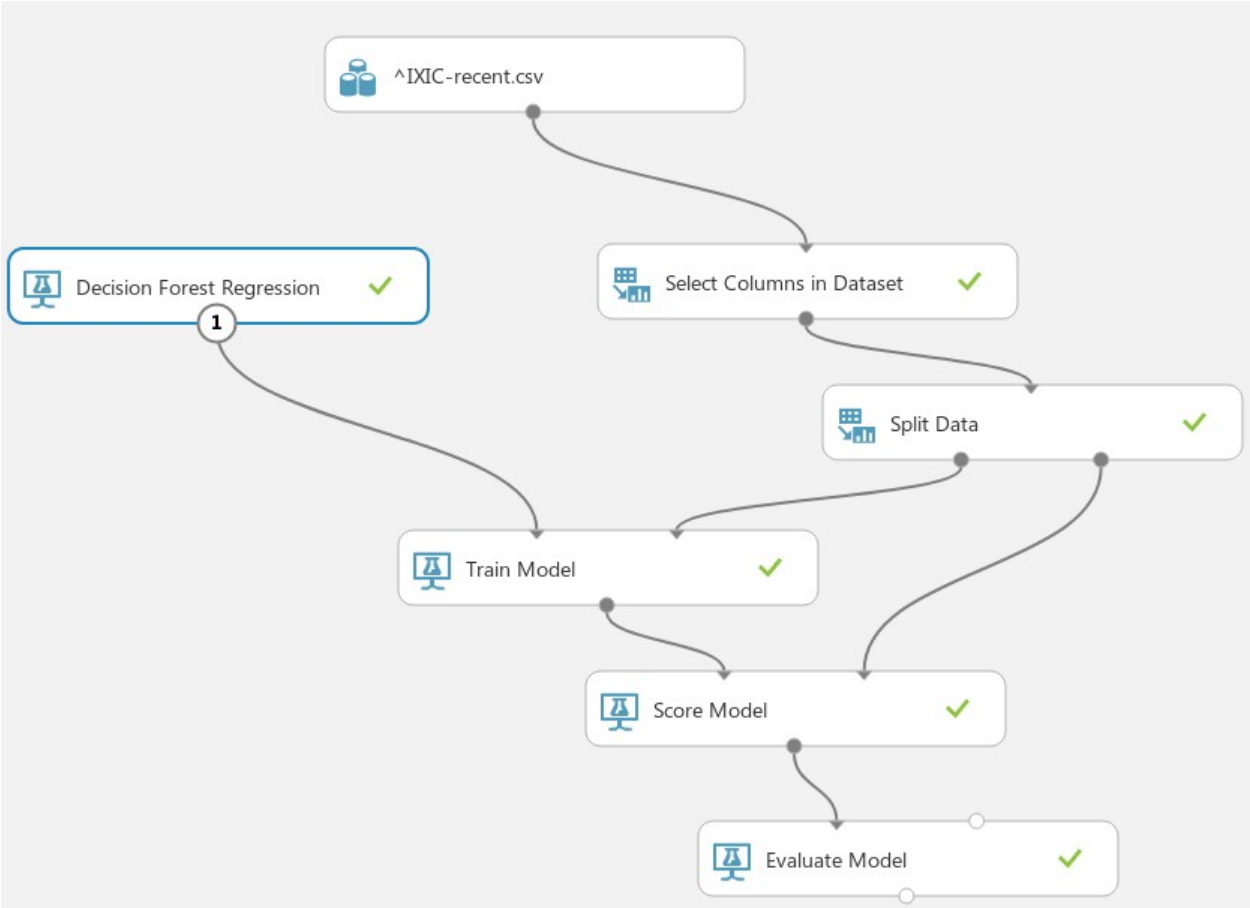
```
import pandas as pd

df = pd.read_csv('^IXIC-recent.csv')
# если в течение дня была получена прибыль, то значение нового столбца 1
# иначе 0
profit = []
for op, cl in zip(df['Open'], df['Close']):
    profit.append(1 if cl > op else 0)
df = df.assign(Profit=profit)
# print(df['Profit'])
df.to_csv("IXIC-redacted.csv")
```

Ссылка на github: https://github.com/memosiki/mai_ai/tree/master/ml1

Decision Forest Regresssion

Оценка модели леса решений для задачи регрессии.

Ссылка на эксперимент: https://gallery.cortanaintelligence.com/Experiment/2-13



Результаты:

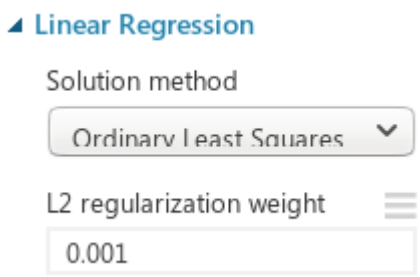
Акции алгоритм 2 › Evaluate Model › Evaluation results

rows	columns
1	6

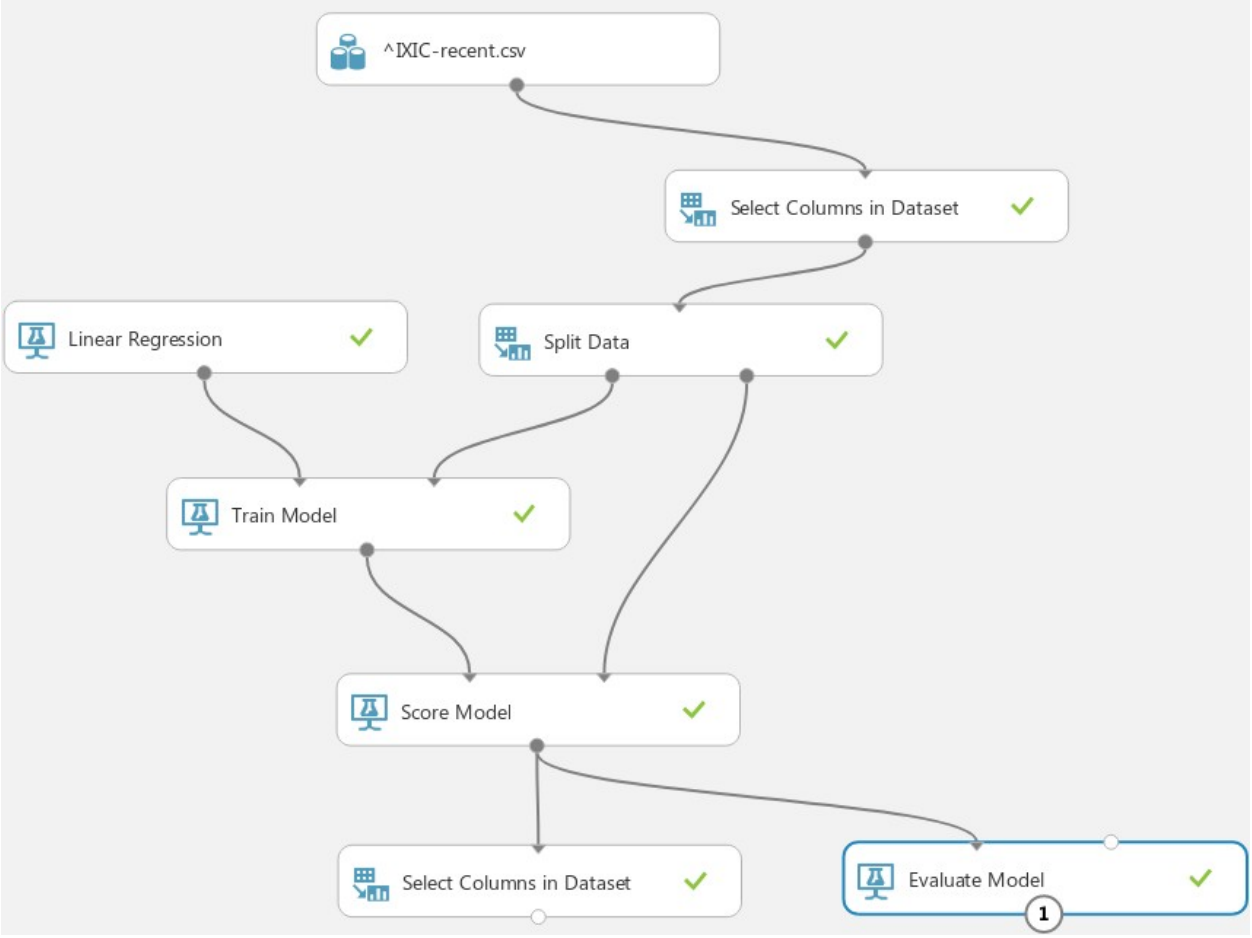
Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
4322.93689	15.253956	23.302455	0.011599	0.000201	0.999799

Алгоритм линейной регрессии

Линейная регрессия исследует зависимость одной переменной от нескольких других переменных с линейной функцией зависимости. Для этого в данной модели используется метод наименьших квадратов.



Ссылка на эксперимент: https://gallery.cortanaintelligence.com/Experiment/1-21



Результаты:

Акции алгоритм 1 › Evaluate Model › Evaluation resu

Metrics

Mean Absolute Error	9.71516
Root Mean Squared Error	14.357516
Relative Absolute Error	0.00717
Relative Squared Error	0.000073
Coefficient of Determination	0.999927

Error Histogram

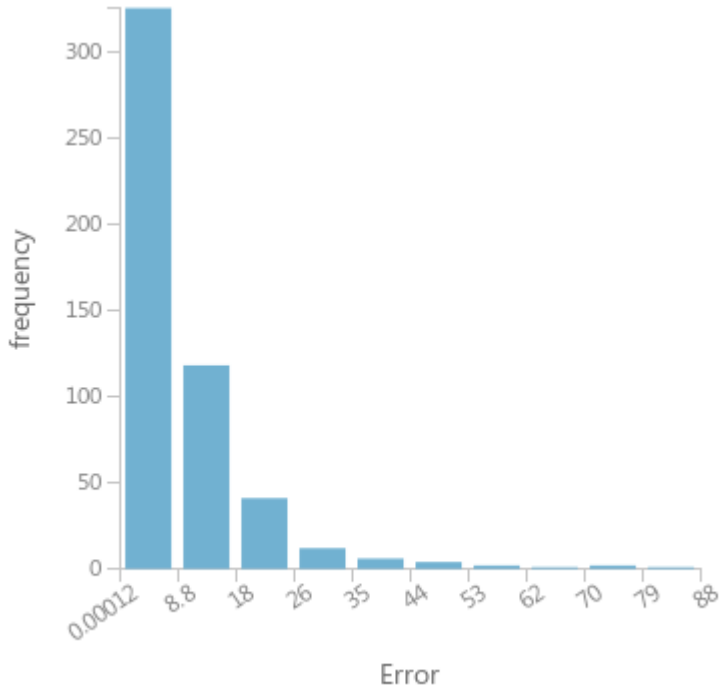
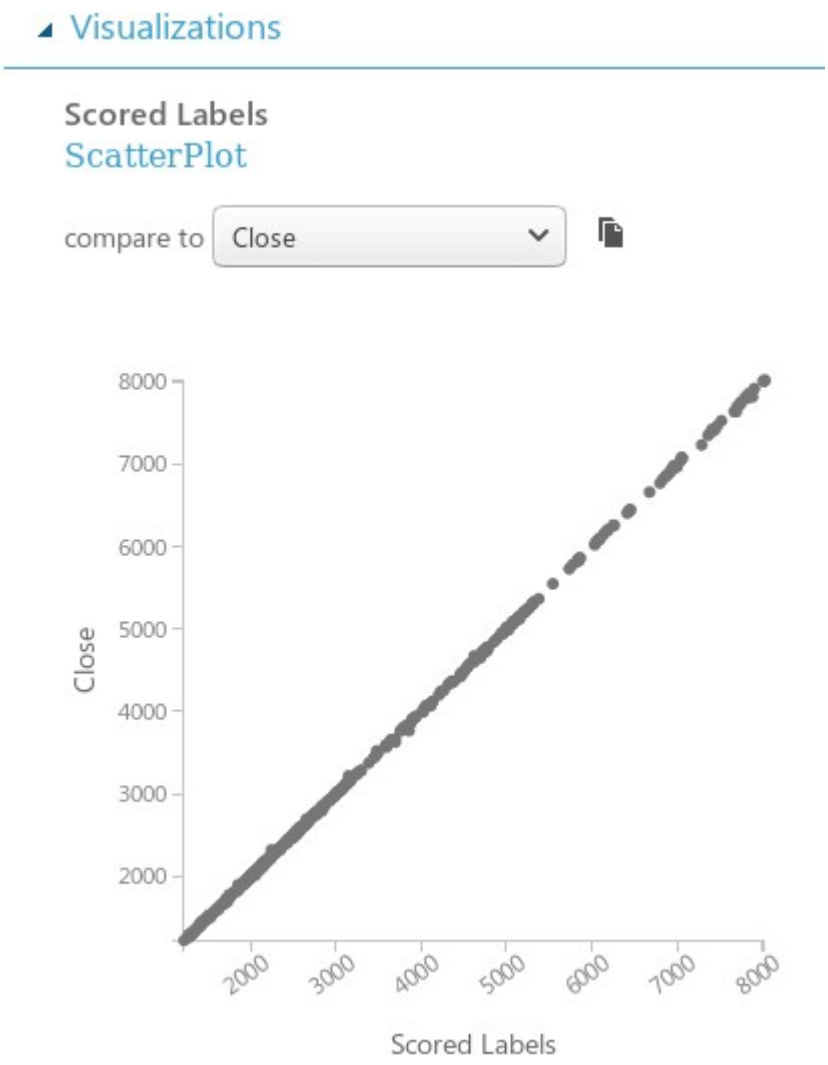


График зависимости полученных значений от ожидаемых:

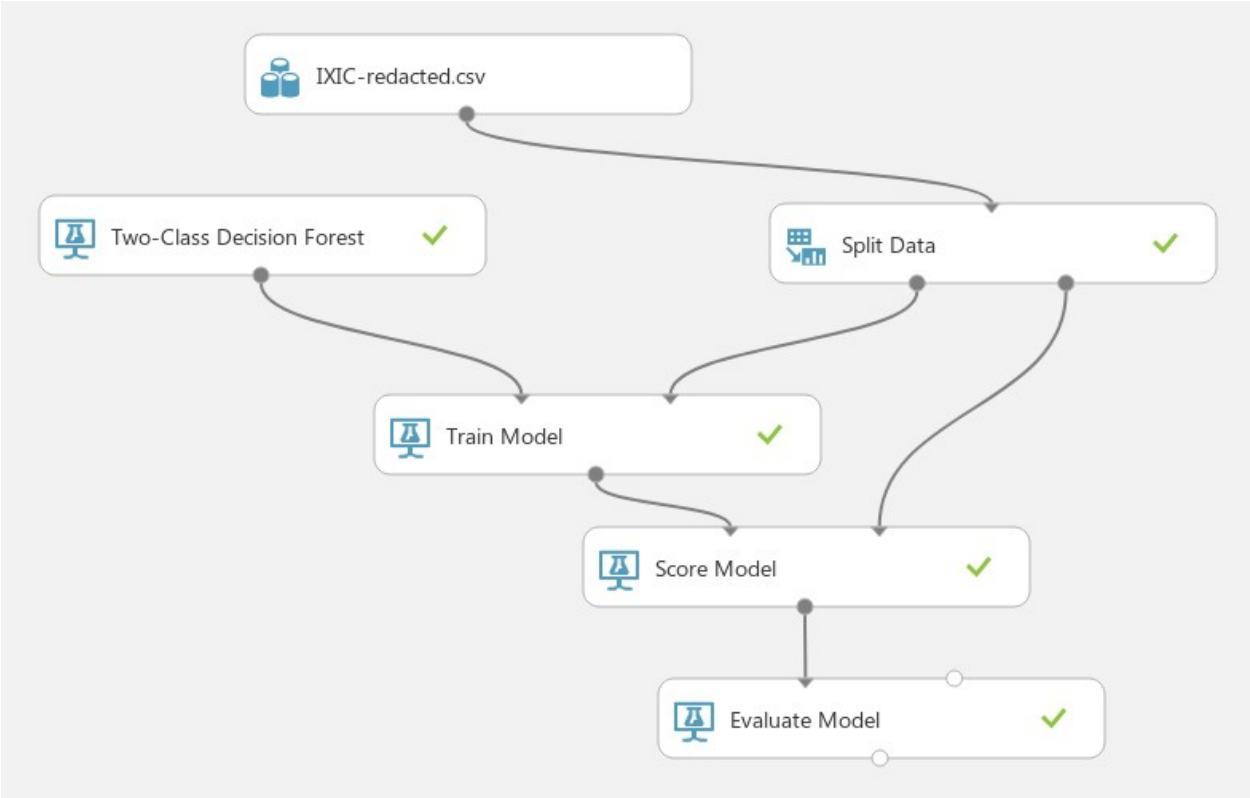


В целом, результаты довольно точны, однако присутствует небольшое количество значительных отклонений. Они связаны с периодами кризиса 2008 года, когда акции могли значительно и неожиданно менять цену в течение дня, естественно данных в таблице недостаточно, чтобы учитывать и их. Лес решений и линейная регрессия показали почти одинаковую точность.

Two-class Decision Forest

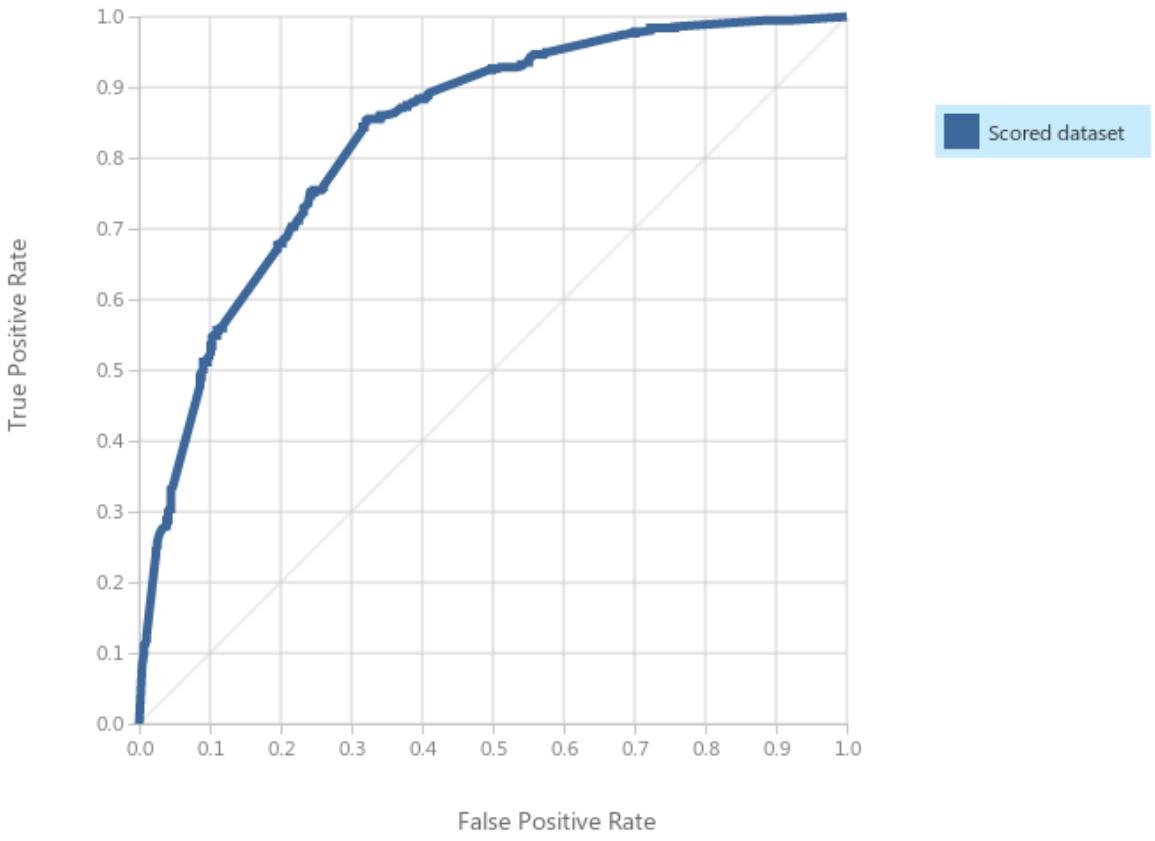
Для того чтобы применить алгоритм классификации к данному датасету введём дополнительный столбец profit, который отображает была ли получена прибыль в данный день торгов.

Применяем лес решений для задачи бинарной классификации


IXIC-redacted.csv
Two-Class Decision Forest
Split Data
Train Model
Score Model
Evaluate Model

Классификация прибилей акций › Evaluate Model › Evaluation results

ROC PRECISION/RECALL LIFT

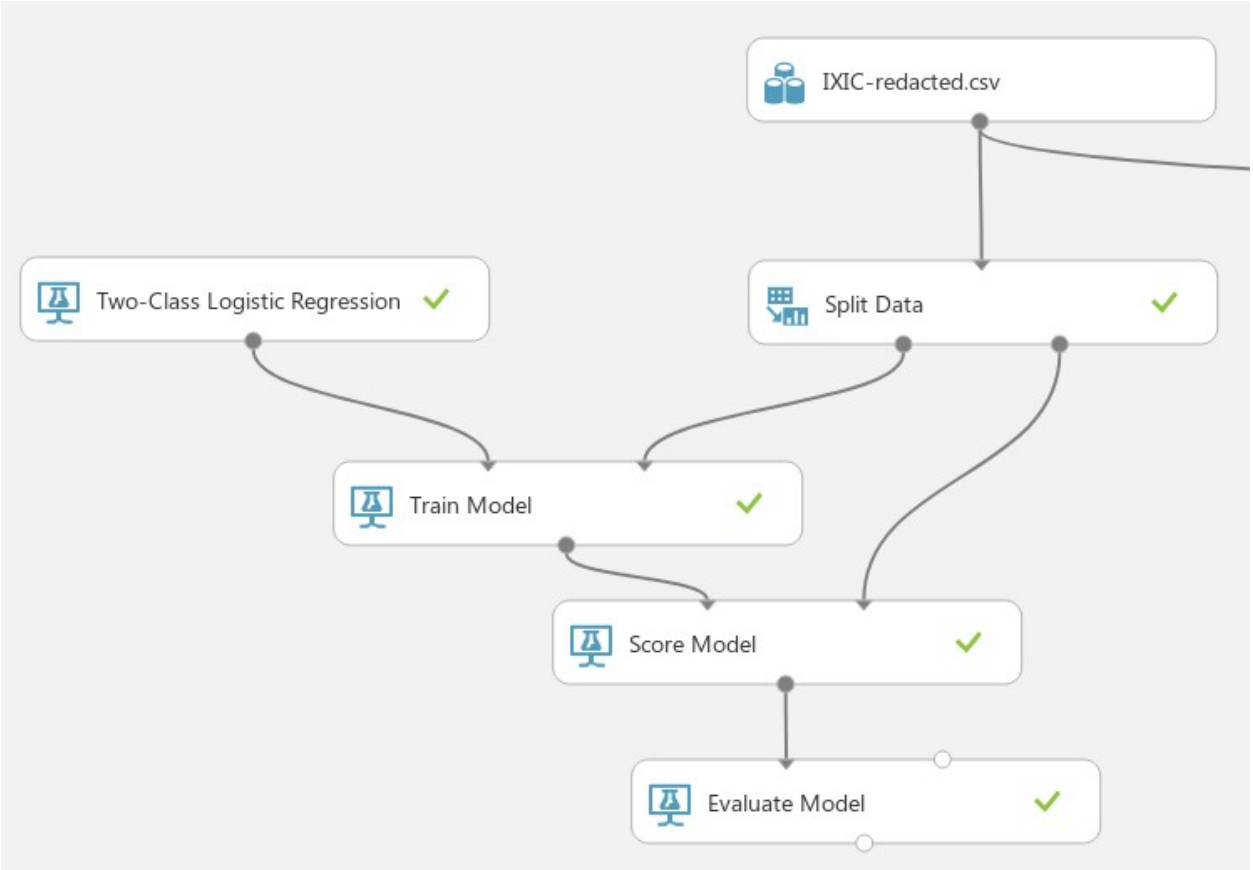


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
425	134	0.751	0.778	0.5	0.833

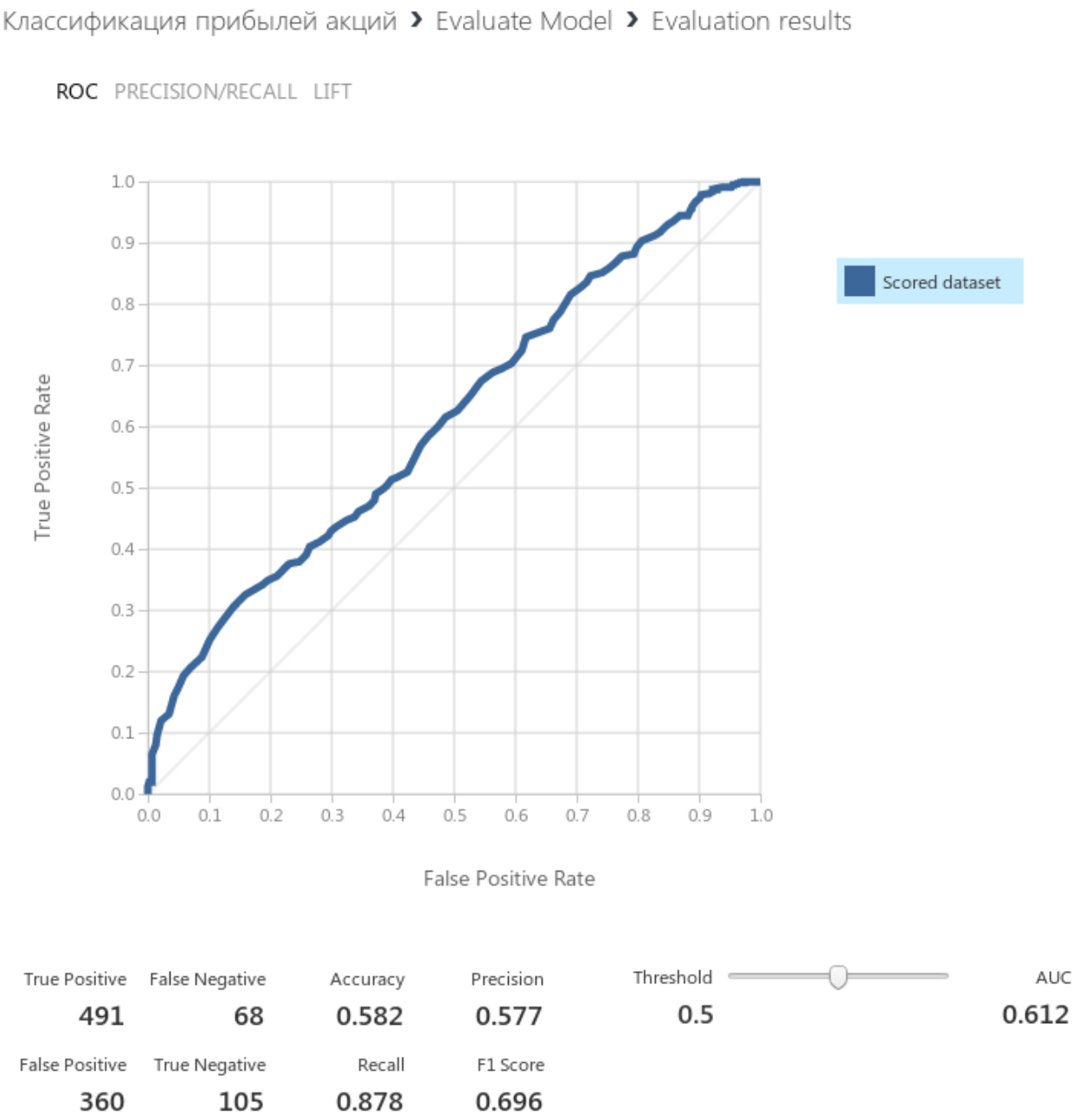
False Positive	True Negative	Recall	F1 Score
121	344	0.760	0.769

Two-class Logistic Regression

Логистическая регрессия используется для прогнозирования вероятности возникновения некоторого события путём подгонки данных к логистической кривой, тем самым логично подходит для классификации данного датасета.


IXIC-redacted.csv
Two-Class Logistic Regression
Split Data
Train Model
Score Model
Evaluate Model

Результаты:



Лес решений показал себя лучше по значению Precision – у него меньше ложных предсказываний.

В тоже время хоть Логистическая регрессия и показала себя в целом хуже, однако по параметру Recall, даже немного превзошла результат леса решений – она лучше определяет положительные ответы.

Ссылка на эксперимент:
https://gallery.cortanaintelligence.com/Experiment/4bd466b2fcd043e898e6804375fbf5a2

Выводы.

Выполнив лабораторную работу, я ознакомился с Microsoft Azure Machine Learning Studio. Я был приятно удивлён что различные алгоритмы машинного обучения очень удобно запускать в облаке. Также при работе получаются очень наглядные модели.