УДК 004.852

Исследование возможности прогнозирования нахождения студентов вуза в зоне риска отчисления на задачах бинарной классификации

Берсенев Александр Иванович
Ярославский государственный технический университет
Студент

Яненко Артем Юрьевич
Ярославский государственный технический университет
Студент

Бойков Сергей Юрьевич
Ярославский государственный технический университет
Канд. тех. наук, доцент

Аннотация
В статье рассматривается система, которая на основе предоставляемого набора данных прогнозирует, что попадет ли конкретный студент из списка в зону риска на отчисления.
Ключевые слова: Машинное обучение, Microsoft Azure Machine Learning Studio

Study of the possibility of predicting the location of university students in the risk zone of deductions on the problems of binary classification

Bersenev Aleksandr Ivanovich
Yaroslavl State Technical University
Student

Yanenko Artem Yurevich
Yaroslavl State Technical University
Student

Bojkov Sergej Yurevich
Yaroslavl State Technical University
candidate of technical sciences, Associate Professor.

Abstract
The article discusses a system that, based on the data set provided, predicts whether a particular student will fall into the risk zone for deductions from the list.
Keywords: Machine Learning, Microsoft Azure Machine Learning Studio

В настоящее время искусственный интеллект проникает во все аспекты жизни общества. Одним из подразделов искусственного интеллекта, изучающих построения алгоритмов способных к обучению является машинное обучение. Простым примером задачи для машинного обучения является задача классификации, когда на основе входных данных, объекту присвоиться какой-либо класс. Для примера разработаем систему, которая на основе предоставляемого набора данных прогнозирует попадет ли конкретный студент из списка в зону риска на отчисления. Для ВУЗа эта система позволит сократить финансовые потери в связи с отчислением студентов. С помощью такой системы ВУЗ будет иметь представление о том, какие студенты находятся на грани отчисления, и сможет принять более точное решение о дальнейшем будущем студента. Это позволит сократить кол-во человеко-часов, потраченных на бесперспективных студентов, а также поспособствует выявлению проблем, связанных с неуспеваемостью.

Данная задача, в машинном обучении называется задачей бинарной классификации. Она состоит в определении к какому классу из двух изначально известных относится данный объект. Также имеются образцы каждого класса — объекты, про которые заранее известно к какому классу они принадлежат. Такие задачи называют обучением с учителем, а известные данные называются обучающей выборкой [1].

Набор данных состоит из следующих полей: Данные об абитуриенте(пол, результаты ЕГЭ, средних балл при поступлении по аттестату, место жительства, данные о предыдущем образовании, год окончания школы, данные об индивидуальных достижениях, особые права при поступлении), данные обучении студента в Вузе(Направление подготовки, факультет, Живет ли в общежитии, Бюджет или платно, данные об успеваемости, Данные о посещении вуза на основе СКУД, отчислен/не отчислен).Для разработки и обучения нейронной сети, было решено использовать Azure ML Studio. Перейдем в Azure ML Studio и создадим новый эксперимент. Студия машинного обучения Microsoft Azure— это инструмент для совместной работы, предназначенный для создания, тестирования и развертывания решений для прогнозного анализа данных [2].

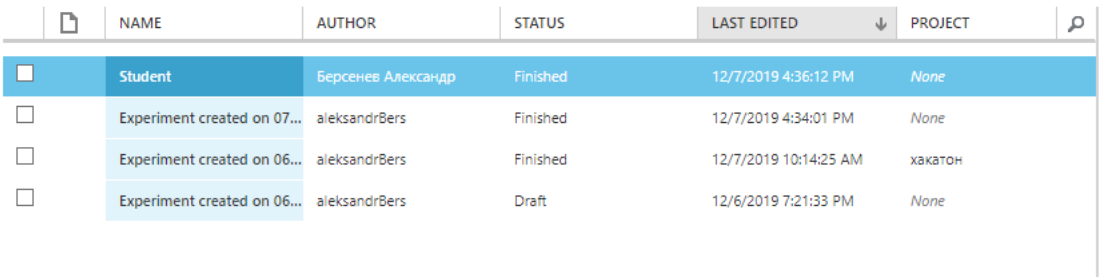

Рисунок 1 – Создание эксперимента в Azure ML Studio

Загрузим в эксперимент наши данные и выберем поля, по которым будет обучаться наша модель.

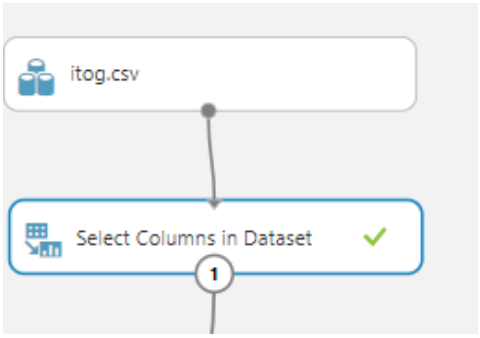

Рисунок 2 – Выбор полей, по которым будет осуществляться прогнозирование

Разделим наши данные на две части. Инструмент Split Data позволяет разделить нашу выборку на две части. Одна из них будет тренировочной выборкой, другая тестовой. Обучающая выборка — выборка, по которой производится настройка модели зависимости. Тестовая — выборка, по которой оценивается качество построенной модели. Если обучающая и тестовая выборки независимы [3].

Рисунок 3 – Разделение выборки на две части

Выберем модель Two-Class Support Vector Machine (Метод опорных векторов) и модель Train Model. Train Model – универсальный компонент, позволяющий обучение любой модели на любой обучающей выборке.

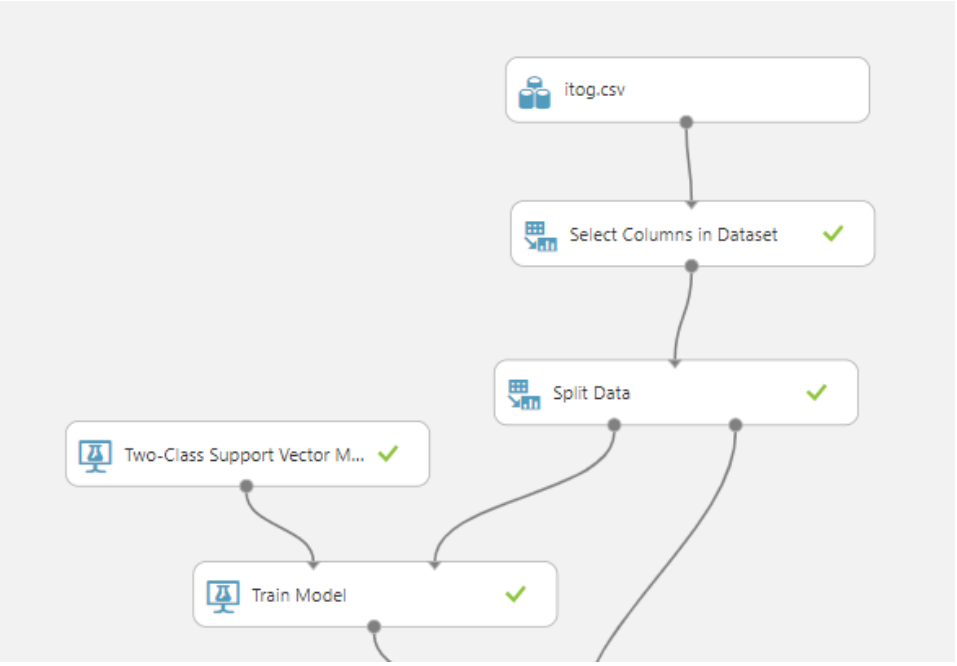

Рисунок 4 – Созданная модель

Идею метода опорных векторов можно показать на простом примере: точки в пространстве разбиваться на два класса. Проводиться линия, разделяющая эти два класса. Далее, все новые точки (не из обучающей выборки) автоматически классифицируются следующим образом: точка над прямой попадает в класс 1, точка под прямой — в класс 2.

Добавим компоненты Score Model и Evaluate Model для того, чтобы рассчитать численные характеристики качества обучения.

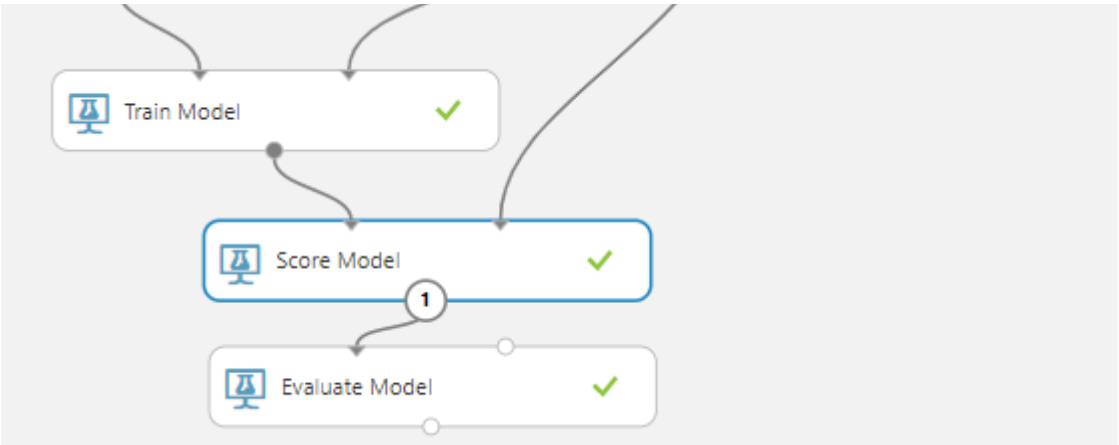

Рисунок 5 – Добавление модулей для отображения результата

Запустим эксперимент и оценим качество нашей модели. Поскольку данные в нашей выборке являются несбалансированными (количество отчисленных студентов меньше, чем количество не отчисленных), показатель Accuracy не будет отражать качество обучения нашей модели. Поэтому для оценки качества будем ориентироваться на отношение ложноотрицательных предсказаний (False Negative) к истиноотрицательным (True Negative).

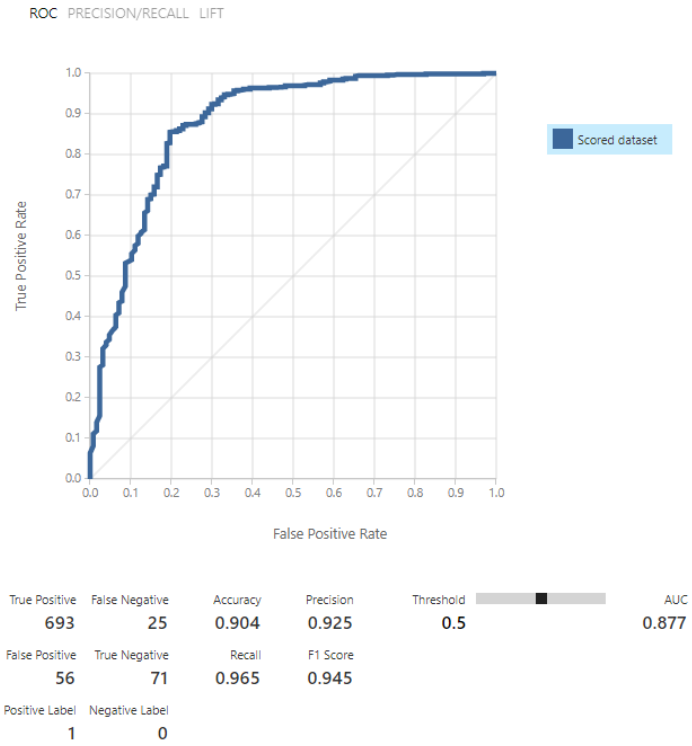

Рисунок 6 – результат Two-Class Support Vector Machine

Как можно видеть модель уже дает неплохой результат. Но для улучшения качества нашей модели, попробуем использовать другие модели обучения вместо Two-Class Support Vector Machine. Далее мы сравним 3 модели для обучения: Two-Class Support Vector Machine, Two-Class Neural Network, Two-Class Logistic Regration.

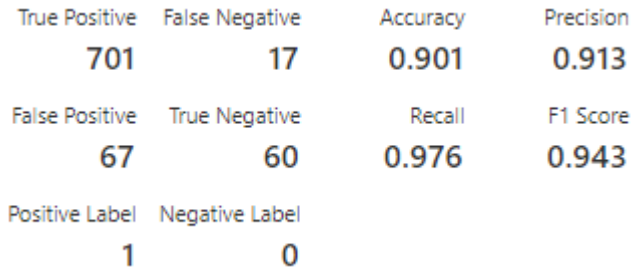

Рисунок 7 – Результат Two-Class Neural Network

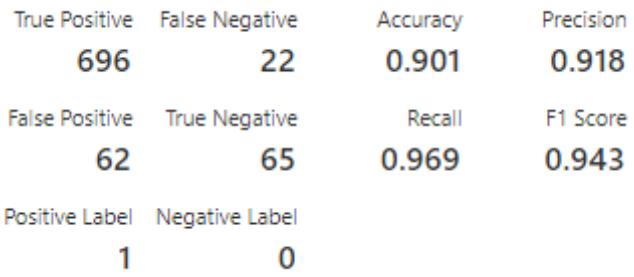

Рисунок 8 – результат Two-Class Logistic Regration

Таблица 1 – сравнение моделей обучения.

Метод	FN/TN
Two-Class Support Vector Machine	0.35
Two-Class Logistic Regration	0.28
Two-Class Neural Network	0.33

Как можно видеть из Таблицы 1, наиболее точно отчисляемых студентов предсказывает Two-Class Support Vector Machine.

В ходе написания стать была разработана модель для прогнозирования студентов, находящихся в зоне риска отчисления.

Библиографический список

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М: Финансы и статистика, 1989.
2. Что такое Студия машинного обучения? // docs.microsoft.com URL: https://docs.microsoft.com/ru-ru/azure/machine-learning/studio/what-is-ml-studio (дата обращения: 04.02.2020).
3. Прикладная статистика: основы моделирования и первичная обработка данных / Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д., М: Финансы и статистика, 1983.