УДК 004.032.26

Барков Денис Владимирович,
магистрант, Ангарский государственный технический университет,
e-mail: barkov.dev@gmail.com
Сенотова Светлана Анатольевна,
к.т.н., доцент, Ангарский государственный технический университет,
e-mail:sveta-senotova@mail.ru

ПРИНЯТИЕ РЕШЕНИЯ КРЕДИТОВАНИЯ ЗАЕМЩИКОВ С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ ПРЯМОГО РАСПРОСТРАНЕНИЯ

Barkov D.V., Senotova S.A.

LENDING DECISION-MAKING TO BORROWERS USING A FEEDFORWARD NEURAL NETWORK

Аннотация. Произведена подготовка исходных данных к моделированию. Подобраны подходящие методы кодирования атрибутивных признаков и методы заполнения пустых значений признаков. Определена эффективность данных методов с помощью статистического анализа. Разработано правило, сокращающее объем исходных данных, основывающееся на статистических свойствах. Разработан класс обработки исходного набора данных.

Ключевые слова: Информационные технологии, наука о данных, нейронные сети.

Abstract. The initial data for modeling has been prepared. Appropriate methods for encoding attribute features and methods for filling in empty values of features are selected. The effectiveness of these methods was determined using statistical analysis. A rule has been developed that reduces the amount of initial data based on statistical properties. A class for processing the initial data set has been developed.

Keywords: Information technology, data science, neural network.

В качестве исходных данных для тестирования и тренировки нейронной сети (НС) были взяты статистические данные американской кредитной компании LendingClubc 2018 по 2020 год. Выбор обоснован тем, что набор данных имеет устойчивую корреляцию с действующей экономической ситуацией в США и отражают влияние инфляции и внешних событий на политику оценки заемщиков.

Перед непосредственной передачей набора данных в НС был произведен его эмпирический анализ, для того чтобы отсеять неинформативные признаки, которые не оказывают весомого влияния на конечный результат классификации. Также были отсеяны признаки, которыми не может располагать заемщик на момент рассмотрения заявки. От общей совокупности был отделен целевой признак «оценка заемщика» и описаны оставшиеся признаки [1].

Дальнейшая обработка данных была реализована с помощью языка программирования Python версии 3.10.2 в качестве отдельного класса.

В первую очередь из csv-файла были считаны исходные данные и подсчитано количество записей и признаков.

Затем происходит попытка преобразования всех строковых литералов в вещественный тип данных, если попытка неудачная, то индексы всех атрибутивных признаков заносятся в отдельный список [1].

Далее кодируются атрибутивные признаки, без учета пустых значений, так как неправильное заполнение или удаление таких признаков, может внести существенные искажения в распределение непрерывной характеристики.

Затем подсчитываются статистические свойства признаков: минимум, максимум, количество уникальных значений, количество пустых значений, среднее значение, медиана, стандартное отклонение и записываются в отдельный csv-файл. Эти свойства необходимы для определения корреляции между признаками, которая позволяет оценить эффективность выбранных методов анализа данных. А также позволяют определить, удаление каких признаков не внесет искажений в выборку данных [1,2].

Основываясь на статистических свойствах, было разработано правило для сокращения объёма выборки. Все признаки, доля пропусков которых более 70%, а также признаки, число уникальных элементов которых более 10000 – удаляются. Такое правило необходимо для того, чтобы избежать сложностей в процессе обучения НС, так как в совокупности, такие выбросы будут только уменьшать корреляцию с целевой переменной, за счет разнородности значений. Из-за большого количества пустых значений признак, зачастую, не имеет никакой нагрузки в узлах НС [2].

В заключении были заполнены пустые значения признаков. Формирование пропусков происходит в зависимости от неизвестных факторов, и информация не может быть восстановлена на основе других атрибутов, из-за того, что возможные взаимосвязанные атрибуты отсутствуют в наборе данных. При таком механизме, игнорирование признаков, имеющих пропуски, приведет к значительному искажению распределения статистических свойств данных.

Для заполнения использовалась медиана каждого признака, так как она более устойчива к выбросам, чем среднее значение. Стоит отметить, что пропуски в атрибутивных признаках также заполнялись медианой, это обосновано тем, что после кодирования признак может быть мультимодален, что свидетельствует о том, что он не подчиняется нормальному закону распределения. Это происходит из-за того, что значения признаков формируются из многих независимых факторов. Следовательно, метод заполнения пропусков медианой не внесет серьёзных искажений или искусственного усиления корреляции [1,2].

ЛИТЕРАТУРА

1. **Барков Д.В., Сенотова С.А.** Кодирование категориальных признаков в нейронных сетях.Сборник научных трудов АнГТУ – Ангарск.:Издательство АнГТУ, 2021, - 411с.

2. **Рашка С.** Python и машинное обучения: / пер. с англ., – М.: ДМК Пресс, 2017, – 418 с.: ил.