**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**

**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**(РУТ (МИИТ)**

**Институт/факультет:** «Академия Высшая Инженерная Школа»

**Кафедра:** «Информатика и вычислительная техника»

**Специальность/Направление подготовки:** «IT-сервисы и обработка данных на транспорте»

**КУРСОВАЯ РАБОТА**

**По дисциплине:** «Машинное обучение и анализ данных»

|  |  |
| --- | --- |
| **На тему:** | «Прогнозирование появления отказов на локомотивах» |

|  |  |
| --- | --- |
| **Обучающийся:** Орлов Вадим Михайлович | **( )** |
| **Проверил:** | **( )** |

**Москва 2022 г.**

**ВЫПОЛНЕННАЯ РАБОТА**

Содержание:

1. Загрузка массива целей
   1. Загрузка данных в программу
   2. Формирование массива параметров
2. Объединение массива целей с массивом признаков
   1. Слияние массивов по ключам
3. Обработка полей после объединения
   1. Удаление и замена пропущенных значений
4. Выбор модели и метода ее обучения
5. Преобразования данных
6. Инженерия данных
   1. Разделение массива на тестовую и тренировочную выборки
   2. Нормализация данных
   3. Балансировка данных
7. Обучение модели
8. Проверка качества обученной модели
   1. Оценка встроенными средствами использованных библиотек
   2. Оценка методами, построенными лично

**1. Загрузка массива данных**

Сперва были загружены массивы целей и признаков, сделана выборка из массива признаков по заданному ключу – *fail\_reason = “Личи”*, удалены столбцы [*fail\_place, fail\_reason, fail\_type*], все столбцы обоих массивов были переведены в нижний регистр, поля с датами были преобразованы в соответствующий формат. Для продолжения работы, было необходимо объединить оба набора данных в один массив, после чего скорректировать – удалить дубликаты, убрать пустые значения.

**2. Объединение массива целей с массивом признаков**

Данные были объединены в один общий массив признаков и целей по ключам [*ser\_loc, znc\_loc, prs\_loc*].

**3. Обработка полей после объединения**

В результате объединения массивов возникло множество ячеек с пустыми значениями, незначимых столбцов. Все это было удалено в несколько этапов:

1. Были удалены поля, где значение в столбце *[count\_fail] = NaN*
2. Были удалены столбцы *[date\_repair\_finish\_y\_1, date\_repair\_finish\_x\_2, date\_repair\_finish\_y\_2, date\_repair\_finish, date\_repair\_finish\_x\_1]*

**4. Выбор модели и метода ее обучения**

В качестве предсказательной модели был выбран классификатор, разработанный компанией «Яндекс» - CatBoost. Был создана модель для ее будущего обучения.

**5. Преобразования данных**

Значения в столбце day были приведены к следующему виду:

* Значения NaN были заменены на 0
* Значения, отличные от нуля и не равные NaN, были заменены на 1

Кроме того, использованные для объединения ключи [*ser\_loc, znc\_loc, prs\_loc*] были назначены индексами объединенного массива.

**6. Инженерия данных**

6.1. Разделение массива на тестовую и тренировочную выборки.

Объединенный массив данных необходимо разделить на две большие части – часть, именуемую X, содержащую параметры, по которым модель будет предсказывать некие данные, и на часть, называемую У, содержащую данные, соответствующие определенным параметрам из части Х. При обучении моделей, разработчики исходят из того, что данные из массива Х неким образом связаны с массивом У, что конкретизирует задачу создаваемой модели – модель должна найти взаимосвязь между параметрами Х и данными У. Однако, просто отдать модели разделенные данные нельзя – мы просто не сможем проверить, как она работает. Для этого необходимо разделить полученные выдержки данных Х и У еще раз разделить – на тестовые и тестировочные данные – X\_test, y\_test и X\_train, y\_train соответственно. Пропорция, выбираемая для разделения данных на части определяет количество данных, идущих на обучение модели и на ее тестирование. В нашем случае было выбрано соотношение 4/6, где 4 части отводятся под тесты, а оставшиеся 6 – под тренировку модели.

6.2. Нормализация данных.

Для грамотной работы любой модели последней необходимо предоставить нормализованные ( приведенные к общему конкретному виду ) данные. Для выполнения данной операции был применен метод MinMaxScaler библиотеки sklearn.

6.3. Балансировка данных.

Поскольку в объединенном массиве оказалось очень много нулевых значений У и мало единичных, было принято решение выполнить балансировку данных методом SMOTE. Были сбалансированы тренировочные данные – к тестовым применялась только нормализация.

**7. Обучение модели**

После всех проведенных операций можно приступать к обучению модели классификатора, которому были переданы тренировочные данные для обучения. Была использована функция потерь «Перекрестная Энтропия», пройдена 1000 итераций обучения.

**8. Проверка качества обученной модели**

Для оценки предсказательных способностей модели были использованы следующие методы оценивания: confusion matrix (Матрица ошибок, матрица путаницы) и f1-мера.

Матрица ошибок позволяет наглядно увидеть, сколько было совершено верных предсказаний и сколько – ошибочных, что дает нам понимание того, как в принципе работает наша модель и насколько она хорошо предсказывает данные.

F-оценка или F-мера является мерой точности теста. Она рассчитывается на основе точности и отзыва теста, где точность - это количество истинно положительных результатов, деленное на количество всех положительных результатов, включая те, которые не были идентифицированы правильно, а отзыв - это количество истинно положительных результатов, деленное на количество всех образцов, которые должны были быть идентифицированы как положительный. Точность также известна как положительная прогностическая ценность, а отзыв также известен как чувствительность в диагностической двоичной классификации.