**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №3**

**по дисциплине «Машинное обучение»**

Тема: Исследование алгоритмов классификации

|  |  |  |
| --- | --- | --- |
| Студент гр. 2372 |  | Юрин А. А. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Знакомство с задачей классификации, изучение методов k-ближайших соседей и дерева решений.

**Постановка задачи.**

1. Выбрать атрибут, выступающий в роли целевого класса (метки);
2. Оценить сбалансированность классов;
3. Применить методы классификации алгоритмами kNN, Дерево решений;
4. Оценить и сравнить результаты метриками оценки качества accuracy, precision, recall, f1, ROC;
5. Проанализировать полученные результаты.

**Выполнение работы.**

Был выбран уже использовавшийся в прошлой лабораторной работе набор данных по прогнозированию отмен бронирования отелей. Набор данных был сначала обработан, были заменены категориальные признаки на числовые.

Набор данных предназначен для изучения классификации и регрессии. Целевым классом для классификации является booking\_status – атрибут, отвечающий на вопрос «Было ли отменено бронирование?».

При оценке сбалансированности этого класса оказалось, что соотношение неотменённых бронирований к отменённым бронированиям составляет 24390 к 11885, соответственно. Было принято решение провести балансировку методом over-sampling, а также проверить результаты для несбалансированного набора данных.

Для двух полученных наборов данных был применён kNN алгоритм (k-ближайших соседей) с выбором k методом плеча. Также была обучена модель Decision Tree (дерево решений).

Результаты обучения отображены на рисунке 1.

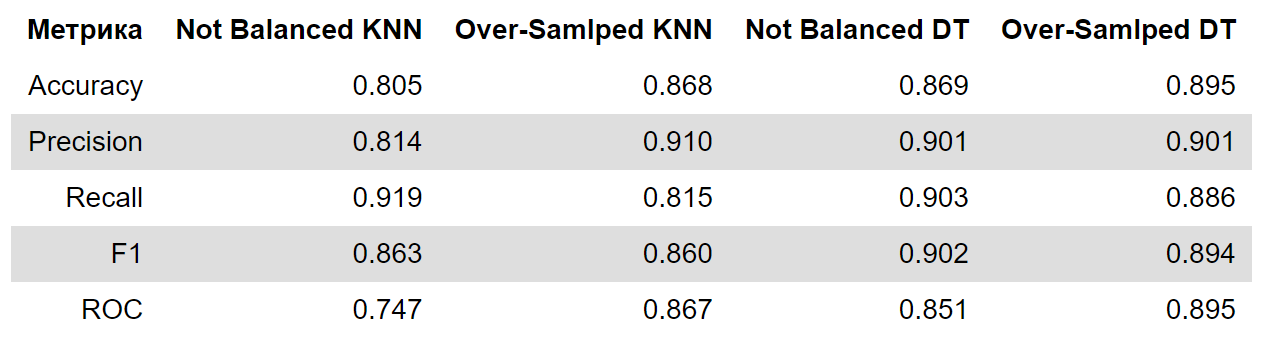


Рисунок 1. Результаты классификации.

На основе полученных результатов метрик можно сделать выводы:

1. Модели на основе DT показывают более высокие результаты, чем KNN, как при использовании оригинальных данных, так и с применением метода over-sampling. Особенно это заметно по метрике accuracy, где DT достигает 0.895 против 0.868 у KNN. Это также подтверждается улучшенными значениями по меткам Precision, Recall и ROC, что делает деревья решений более подходящими для данной задачи.
2. Использование метода балансировки over-sampling значимо улучшает результаты для обеих моделей. Особенно это видно по Accuracy и F1. Модели, обученные на сбалансированных данных, показали наилучшие результаты, подтверждая, что переподборка классов помогает справиться с дисбалансом данных, улучшая общую производительность.
3. Несмотря на высокие показатели Recall у KNN без балансировки (0.919), модель имеет наименьшую Accuracy (0.805) и ROC (0.747). Это указывает на проблему переобучения или переоценки одних классов в ущерб другим. В то время как KNN в целом может показывать высокую полноту, точность модели страдает, и она недостаточно хорошо разделяет классы.
4. Модели на основе DT, особенно с применением переподборки классов, показывают лучший баланс между Precision, Recall, и F1. Это делает их более универсальными для различных типов задач, где важны не только точность, но и полнота классификации.

**Выводы.**

В ходе работы были изучены алгоритмы классификации KNN и дерево решений. Проведено обучение моделей, оценка и сравнение результатов.