Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления» (ИУ)

КАФЕДРА «Системы обработки информации и управления» (ИУ5)

КУРСОВАЯ РАБОТА ПО КУРСУ ТЕХНОЛОГИЯ МАШИННОГО ОБУЧЕНИЯ

Студент:
Мурзин В.В., группа ИУ5Ц-81Б _____ (подпись, дата)

Преподаватель:
Гапанюк Ю.Е. _____ (подпись, дата)

Москва, 2021

СОДЕРЖАНИЕ

Введение

Умение проводить разведочный анализ данных и подготавливать данные для последующего анализа с использованием методов машинного обучения необходимо для решения большинства прикладных задач, связанных с использованием методов машинного обучения, в частности нейронных сетей, и в целом с анализом данных, моделированием процессов и прогнозированием.

Эта способность применять правильные подходы к решению комплексных задач машинного обучения, а также умение обосновывать свои решения на каждом этапе работы, является необходимым требованием для любого технического специалиста.

Основная часть

1.1 Постановка задачи

Целью данной работы является решение комплексной задачи машинного обучения и обоснование принятых решений.

Для достижения цели данной работы были решены следующие **задачи**:

- Поиск и выбор набора данных для построения моделей машинного обучения
- Формирование тестовой и тренировочной выборок
- Проведение разведочного анализа данных
- Выбор признаков подходящих для построения моделей
- Проведение корреляционного анализа данных
- Выбор метрик для последующей оценки качества моделей
- Выбор наиболее подходящих моделей для решения задачи классификации
- Построение базового решения
- Подбор гиперпараметров для выбранных моделей
- Сравнения качества полученных моделей с базовым решением
- Формирование выводов о качестве моделей на основе полученных результатов

1.2 Алгоритм решения

1.2.1 Поиск и выбор набора данных

В рамках этого проекта была решена **задача классификации** того, является ли пойманное насекомое переносчиком вируса Западного Нила (*West Nile virus, WNV*) на основе данных [1] с *Kaggle.com*, открытой платформы для проведения соревнований по машинному обучению.

Решаемая задача ***актуальна***, поскольку анализ геоинформации и результатов тестов во времени позволяет проследить закономерности появления групп зараженных насекомых.

Данные представляют собой **два** набора: *train.csv*, *test.csv* - обучающий и тестовый наборы данных.

Обучающий набор состоит из данных за **2007**, **2009**, **2011** и **2013** годы, а в тестовом наборе предлагается предсказать результаты тестов за **2008**, **2010**, **2012** и **2014** годы.

Наборы данных содержат следующие поля признаков:

1. *Id*: идентификатор записи
2. *Date*: дата проведения теста на *WNV*

3. *Address*: приблизительный адрес местонахождения ловушки. Используется для отправки в геокодер.
4. *Species*: вид комаров
5. *Block*: номер квартала по адресу
6. *Street*: название улицы
7. *Trap*: идентификатор ловушки
8. *AddressNumberAndStreet*: приблизительный адрес, полученный от геокодера
9. *Longitude, Latitude*: широта и долгота, полученные от геокодера
10. *AddressAccuracy*: точность, полученная от геокодера
11. *NumMosquitos*: количество комаров, пойманных в эту ловушку
12. *WnvPresent*: присутствовал ли вирус Западного Нила в этих комарах. **1** означает, что *WNV* присутствует, а **0** - нет.

Целевой признак *WnvPresent* присутствует только в обучающем наборе. Задача состоит в том, чтобы получить предсказания для этого признака для тестового набора данных.

Это **задача бинарной классификации** на несбалансированной выборке.

Поскольку тестовые данные (*test.csv*) не размечены, будут использоваться только обучающий набор (*train.csv*).

Таким образом, в качестве тестовых данных будут использоваться данные за **2013** год. А в качестве обучающего данные за **2007**, **2009**, **2011**.

1.2.2 Выбор признаков и подготовка данных

Признаки в этом наборе данных в основном категориальные, не включая широту, долготу и количество насекомых, поэтому основные усилия по подготовке данных были направлены на работу с категориальными переменными.

Проводился анализ категориального соответствия между обучающей и тестовой выборкой отдельно (оба набора всегда анализировались отдельно, поскольку тестовый набор представляет собой будущий временной интервал). Категориальные переменные были закодированы на основе этого анализа каждой переменной. Где-то количество переменных было уменьшено, чтобы полностью перекрывать друг друга в обучающей и тестовой выборке.

В конечном итоге, все категориальные переменные были закодированы на основе обучающей выборки, используя простой метод последовательного кодирования.

Обучающая выборка была сбалансирована методом ***S**ynthetic **M**inority **O**versampling **TE**chnique* [2] (*SMOTE*) - методом техничного увеличения выборки по классам.

Распределение обучающей выборки по классам см. в приложении (Рис. 1.П.).

1.2.3 Корреляционный анализ подготовленных данных

Исходя из полученной корреляционной матрицы (Рис. 1), был сделан вывод, что наиболее коррелирующими признаками с целевой переменной являются признаки сезона и количества насекомых, протестированных в определенном месте.

Не радует тот факт, что количество насекомых не только сильно коррелирует с целевой переменной, но и изменяется положительно, что может означать, что в ходе исследования была допущена ошибка, и что шанс обнаружения вируса в образце в конкретном месте больше зависит не от места (что необходимо изучить, и изучается в исследовании), а от количества протестированных насекомых.

Тем не менее был сделан вывод, что данные подходят для дальнейшего анализа с помощью **нелинейных методов машинного обучения**.

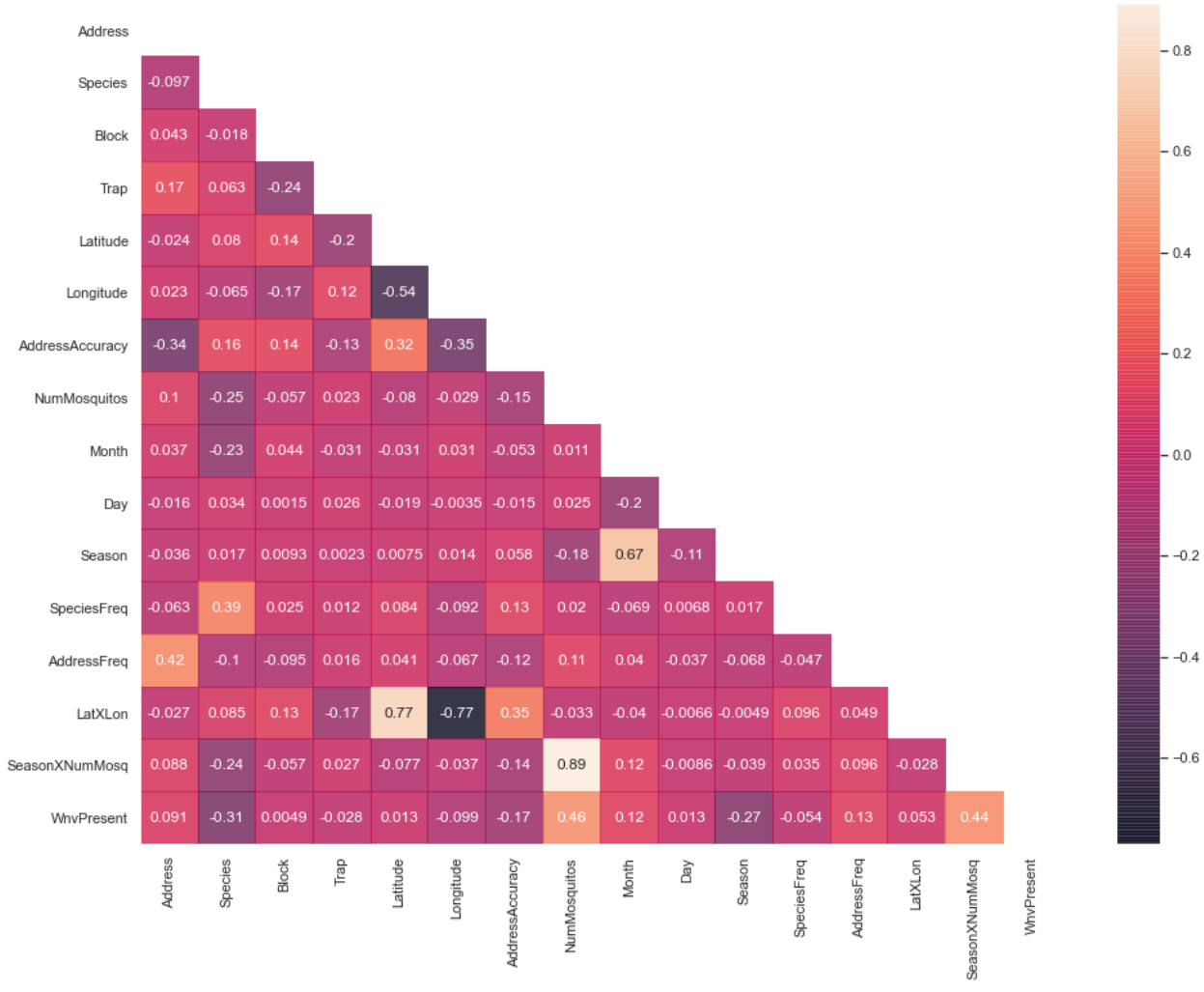


Рисунок 1. Матрица корреляции по методу *Kendall*, более подходящего для категориальных

переменных

1.2.4 Использованные метрики качества

Для оценки качества бинарной классификации на несбалансированной выборке стоит использовать метрики полноты, специфичности и их среднее гармоническое:

Специфичность (*precision*):

Отвечает на вопрос, сколько ресурсов было потрачено:

$$precision = \frac{TP}{TP + FP}$$

Полнота (*recall*):

Отвечает на вопрос, сколько объектов было пропущено:

$$recall = \frac{TP}{TP + FN}$$

Гармоническое среднее полноты и специфичности (*F1*):

Среднее между затраченными ресурсами и пропущенными объектами:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Тем не менее поскольку обучающая выборка была сбалансирована методом *oversampling*, для наглядности была включена в набор метрик и обычная метрика точности.

Более того, исследователи, собравшие набор данных, используемый в данной работе, предложили использовать метрику *ROC-AUC* [3], которая показывает, насколько хорошо классы разделяются друг от друга при предсказании.

Поэтому в окончательном сравнении результатов были использованы метрики *F1* и *ROC-AUC*.

1.2.5 Методы машинного обучения выбранные для решения задачи

Для решения выбранной задачи классификации были использованы следующие методы машинного обучения:

- Дерево решений
- Лес решающих деревьев
- Бустинг над решающими деревьями
- Мета алгоритм стекинга с логистической регрессией, на лесе решений
- Мета алгоритм стекинга с логистической регрессией, на бустинге над решающими деревьями

Приведенные выше методы были выбраны потому, что дерево решений, лежащее в основе этих подходов, хорошо работает с категориальными переменными и составляет основу многих ансамблевых методов.

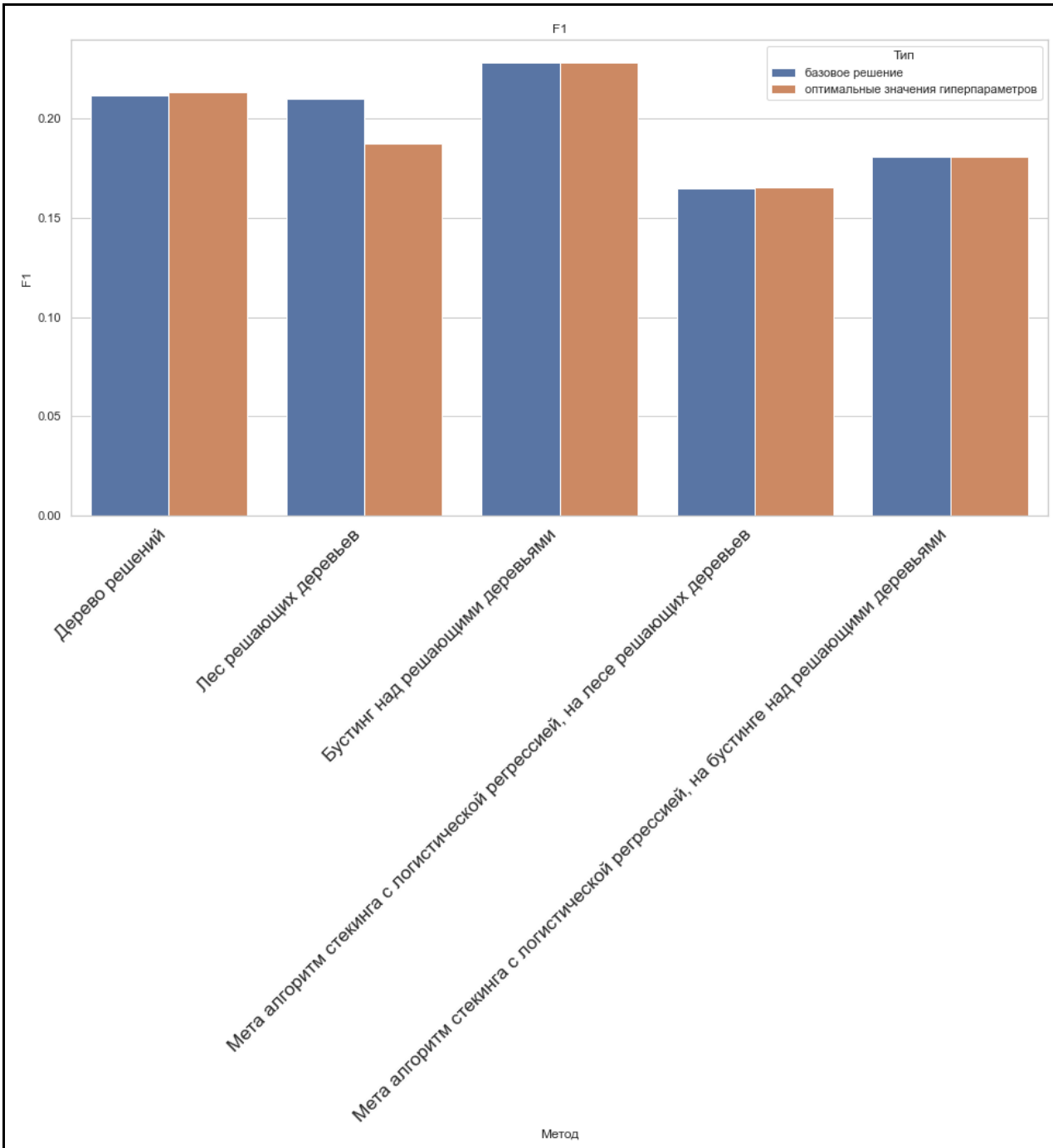
Заключение

В ходе решения задачи классификации были выбраны и протестированы пять алгоритмов машинного обучения, подходящих для классификации данных с большим количеством категориальных переменных. Оптимальные гиперпараметры подбирались с помощью метода поиска по сетке с перекрестной валидацией.

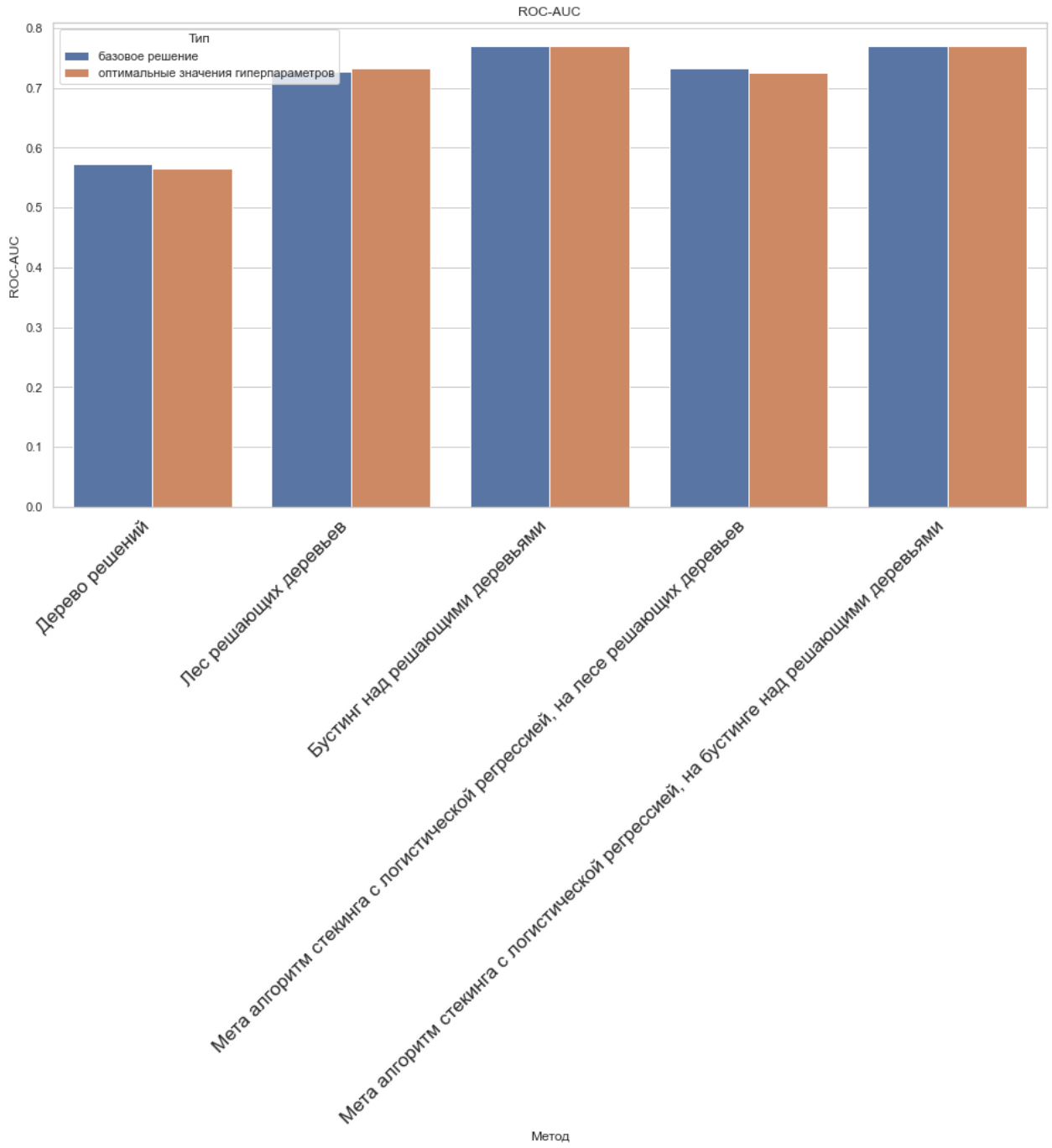
Результаты позволили сделать вывод, что наилучшим способом решения задачи является алгоритм - бустинга над решающими деревьями, реализованный в библиотеке LightGBM [4].

Тем не менее, наиболее интерпретируемые результаты при относительно высоком значении метрики полноты на тестовой выборке - могут быть получены и методом классификации с помощью дерева решений. Поскольку задача носит прикладной характер, интерпретируемость дерева решений (Рис. 3) может помочь исследователям в мониторинге распространения вируса.

Были получены следующие результаты:



F1 (*только 0.03-0.09% всех наблюдений являются положительными)



ROC-AUC

Рисунок 2. Сравнение результатов классификации до и после подбора оптимальных гиперпараметров

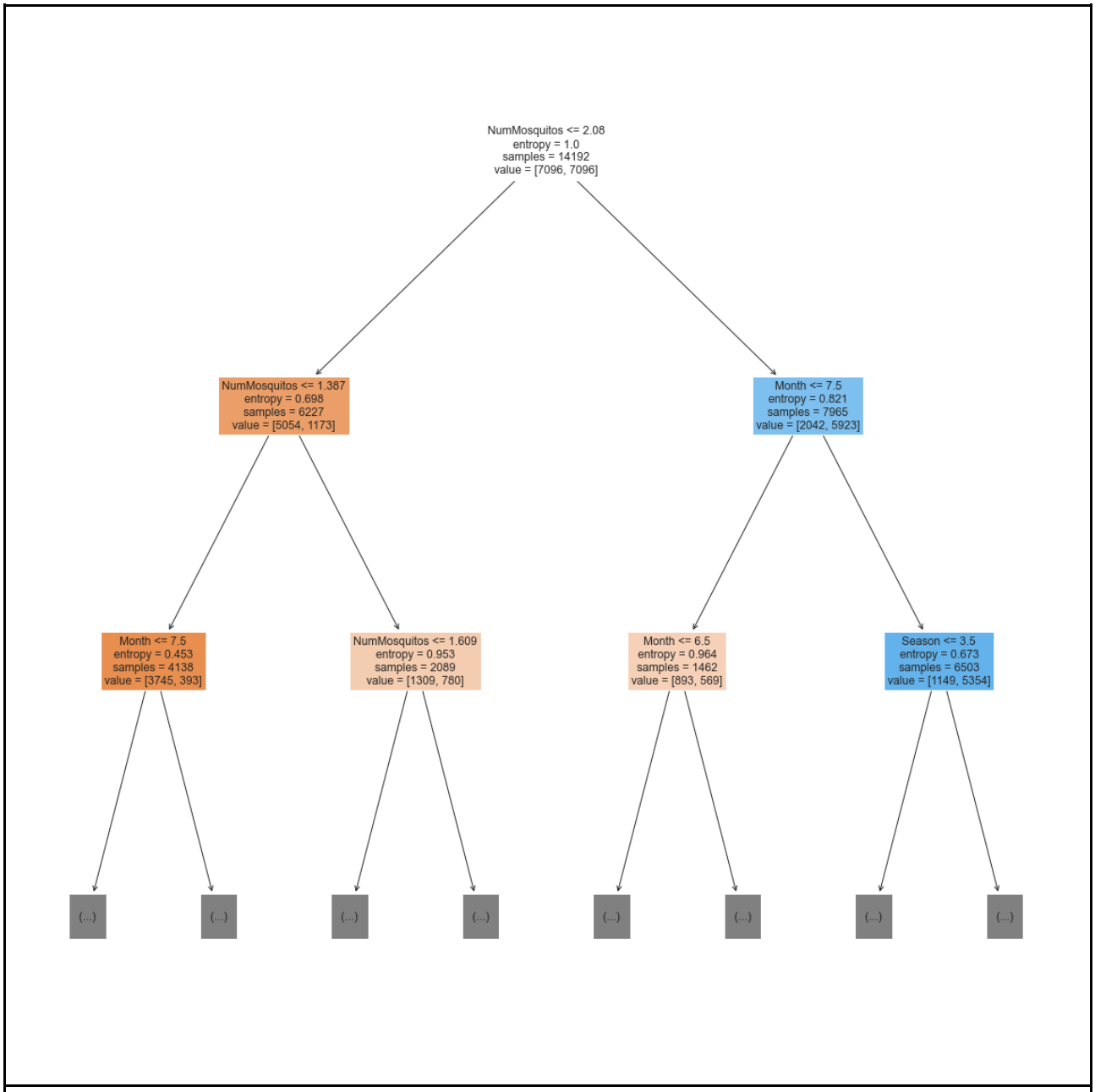


Рисунок 3. Визуализация работы дерева решений

Таким образом, наблюдая за работой дерева решений с оптимальными параметрами, можно еще раз убедиться, что от числа насекомых в исследуемой выборке зависит вероятность нахождения в этой выборке насекомого-носителя вируса. При этом количество комаров зависит от времени года. И созданная в данном проекте дополнительная переменная - сезон, а также месяц - хорошо показывают эту зависимость.

Список использованной литературы

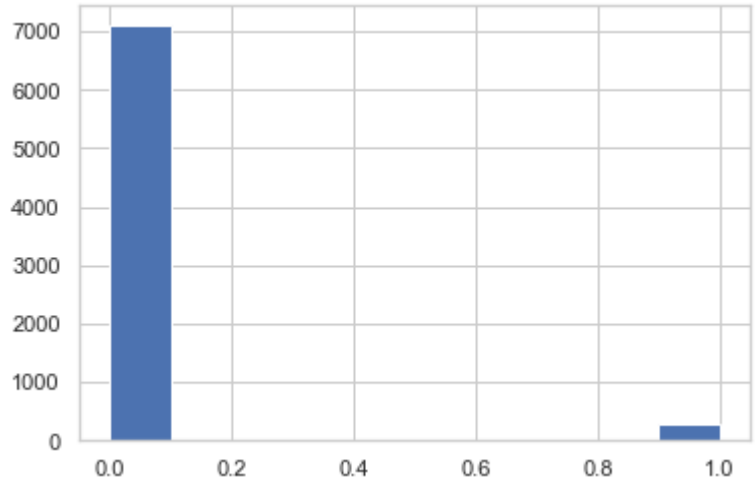
1. predict-west-nile-virus // Kaggle.com [Электронный ресурс]. URL: https://www.kaggle.com/c/predict-west-nile-virus/overview.

2. Chawla N., Bowyer K. SMOTE: Synthetic Minority Over-sampling Technique 2002. № 18.

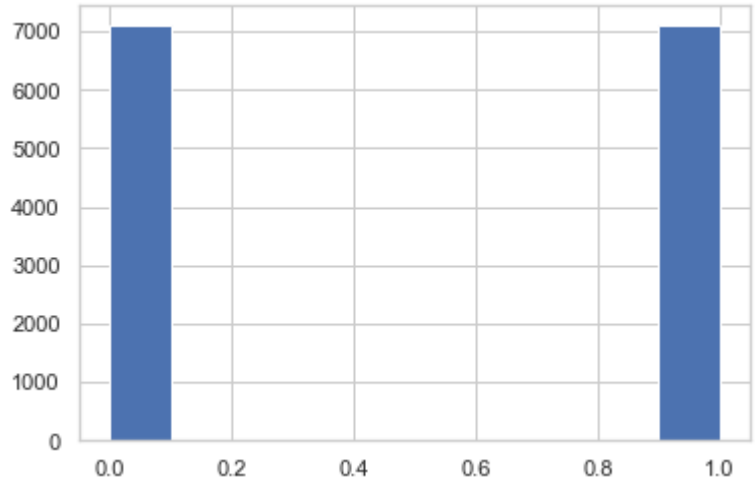
3. Дьяконов А. AUC ROC (площадь под кривой ошибок) // dyakonov.org [Электронный ресурс]. URL: https://dyakonov.org.

4. Ke G. [и др.]. LightGBM: A Highly Efficient Gradient Boosting Decision Tree под ред. I. Guyon [и др.]., Curran Associates, Inc., 2017.

Приложение



Изначальное распределение



Распределение после применения метода *SMOTE-oversampling*

Рисунок 1.П. Распределение по классам до и после применения метода балансировки на обучающей выборке