**УДК 62** **Технические науки**

***Пендюрин Алексей Николаевич,*** *бакалавр 2 курс,*
*Уфимский университет науки и технологий Г. Уфа*
***Заманов Азамат Айратович,*** *бакалавр 2 курс,*
*Уфимский университет науки и технологий Г. Уфа*
***Таушев Филипп Игоревич,*** *бакалавр 2 курс,*
*Уфимский университет науки и технологий Г. Уфа*

# ИССЛЕДОВАНИЕ ЗНАЧИМОСТИ ПРИЗНАКОВ ПЕРЕМЕННЫХ ПРИ ЛОКАЛЬНОМ ДИАГНОСТИРОВАНИИ ПОЛЯРНЫХ СИЯНИЙ

**Аннотация:** В этой статье были рассмотрены различные методы определения значимости признаков в задачах бинарной классификации, такие как метод рекурсивного исключения признаков, метод анализа дисперсии, метод взаимной информации. Метод рекурсивного исключения признаков предоставляет возможность обнаружения наиболее значимого признака в соответствии с выбранной моделью. Метод анализа дисперсии для нахождения различий между средними значениями двух или более групп. Метод взаимной информации для определения общей значимости признаков. При выборе метода определения значимости признаков необходимо учитывать ограничения каждого метода и проводить дополнительные тесты и оценки модели.

**Ключевые слова:** бинарная классификация, признак, метод, модель, полярные сияния, временной ряд, метод рекурсивного исключения признаков, метод анализа дисперсии, метод взаимной информации, информация.

**Abstract:** This article discussed various methods for determining feature importance in binary classification tasks, such as the Recursive Feature Elimination method, the Analysis of Variance method, and the Mutual Information method. The Recursive Feature Elimination method provides the ability to detect the most

significant feature according to the chosen model. The Analysis of Variance method is used to find differences between the means of two or more groups. The Mutual Information method is used to determine the overall importance of features. When choosing a method for determining feature importance, it is necessary to consider the limitations of each method and conduct additional tests and model evaluations.

**Keywords:** binary classification, feature, method, model, auroras, time series, Recursive Feature Elimination method, Analysis of Variance method, Mutual Information method, information.

Введение

В области машинного обучения одной из ключевых задач является отбор признаков, который может существенно повлиять на качество решаемых задач. Отбор признаков включает в себя выбор тех, которые наиболее важны для построения модели. Процесс необходим, поскольку он может помочь улучшить производительность модели, снизить переобучение и улучшить интерпретируемость результатов.

В данной статье мы обсуждаем и анализируем различные методы для оценки значимости признаков в задачах бинарной классификации. Одними из значимых из них являются метод рекурсивного исключения признаков, метод анализа дисперсии и метод взаимной информации.

Метод рекурсивного исключения признаков является эффективным инструментом для отбора наиболее значимых признаков. Метод анализа дисперсии, или ANOVA, помогает определить, какие признаки имеют статистически значимые различия между группами. Наконец, метод взаимной информации применяется для определения важности каждого признака путем вычисления степени взаимной информации между признаком и целевой переменной.

Наша работа фокусируется на исследовании этих методов и их применении на практике, что обеспечивает ценное восприятие в отношении их преимуществ и ограничений, а также их пригодности для различных задач классификации.

Материалы и методы

1. Метод рекурсивного исключения признаков является алгоритмом для выбора наиболее информативных признаков в наборе данных. Он основывается на рекурсивной процедуре, которая последовательно исключает признаки и оценивает их вклад в модель [10].

Алгоритм начинает с построения модели на полном наборе признаков и оценивает их значимость. Затем, признак с наименьшим оценочным показателем удаляется из набора. Процесс рекурсивно повторяется на уменьшенном наборе признаков до тех пор, пока не будет достигнуто определенное условие остановки.

Метод RFE обеспечивает снижение размерности набора признаков, исключая наименее информативные, шумовые или коррелирующие признаки. Он может быть использован для повышения производительности модели, улучшения интерпретируемости и обобщающей способности модели путем фокусировки на существенных признаках.

Этот метод научного исследования является инструментом машинного обучения и широко применяется для выбора признаков в задачах классификации, предоставляя систематический подход к исключению признаков.

В зависимости от задачи, типа данных и выбранной модели, метод рекурсивного исключения признаков может быть применен к любой модели машинного обучения, которая может вычислить значимость признаков. Для большого набора данных с немалым количеством признаков можно использовать модели с линейной зависимостью от признаков. Однако, если признаки нелинейно зависят от целевой переменной, то можно использовать модели, которые учитывают нелинейность.

Для метода рекурсивного исключения признаков были выбраны модели линейной регрессии и случайного леса. Модель линейной регрессии – строится на основе линейной зависимости между зависимой переменной и одной или несколькими независимыми переменными [11]. Модель случайного леса – на

основе объединения нескольких деревьев решений в одну модель. Каждое дерево в случайном лесу строится на основе подмножества, обучающих данных и случайного набора признаков. Так, каждое дерево строится независимо и может оценивать данные по-разному. В результате, модель случайного леса способна улавливать более сложные зависимости между признаками и целевой переменной [12].

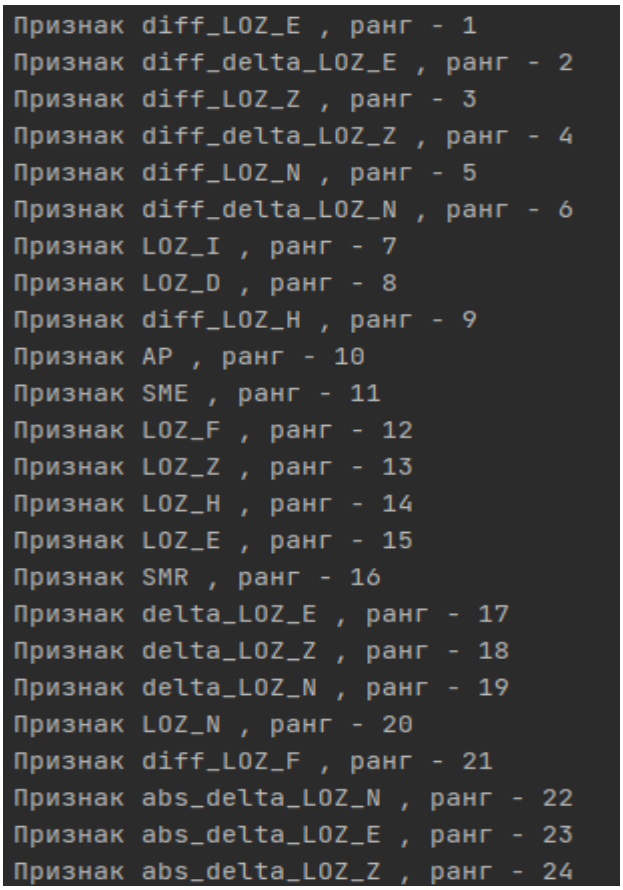


Рисунок 1 – Результаты RFE на основе модели линейной регрессии

```
Признак diff_LOZ_N , ранг - 1
Признак diff_delta_LOZ_N , ранг - 2
Признак LOZ_Z , ранг - 3
Признак delta_LOZ_N , ранг - 4
Признак diff_LOZ_F , ранг - 5
Признак LOZ_E , ранг - 6
Признак LOZ_N , ранг - 7
Признак LOZ_F , ранг - 8
Признак SMR , ранг - 9
Признак LOZ_D , ранг - 10
Признак delta_LOZ_Z , ранг - 11
Признак delta_LOZ_E , ранг - 12
Признак SME , ранг - 13
Признак abs_delta_LOZ_N , ранг - 14
Признак diff_delta_LOZ_Z , ранг - 15
Признак diff_delta_LOZ_E , ранг - 16
Признак diff_LOZ_H , ранг - 17
Признак LOZ_H , ранг - 18
Признак abs_delta_LOZ_E , ранг - 19
Признак abs_delta_LOZ_Z , ранг - 20
Признак diff_LOZ_Z , ранг - 21
Признак AP , ранг - 22
Признак diff_LOZ_E , ранг - 23
Признак LOZ_I , ранг - 24
```

Рисунок 2 – Результаты RFE на основе модели случайного леса

Таким образом, на основе двух моделей для метода рекурсивного исключения признаков, самым значимым является «diff_LOZ_N».

2. Был задействован метод анализа дисперсии ANOVA (Analysis of Variance) - статистический метод, используемый для определения наличия различий между средними значениями двух или более групп. Он также является важным инструментом в области машинного обучения [8].

Метод основывается на сравнении разброса значений данных между группами с дисперсией внутри каждой группы. Если различия между группами статистически значимы, то можно сделать вывод о том, что средние значения в этих группах различаются. ANOVA может использоваться для анализа данных с несколькими факторами, а также для анализа взаимодействия между факторами.

Он широко используется в научных исследованиях, экономике, медицине, психологии и других областях, где требуется оценить значимость различий между группами [9].

Для проведения анализа ANOVA необходимо определить, какие признаки являются факторами, а какие - зависимыми переменными. В данном случае мы искали связь между данными наблюдений и появлением полярных сияний, поэтому "Zenith" являлся зависимой переменной, а остальные признаки - факторами.

| | Feature | F-value | P-value |
|---|---|---|---|
| 2 | AP | 4203.841082 | 0.000000e+00 |
| 19 | diff_LOZ_N | 4202.177327 | 0.000000e+00 |
| 16 | diff_delta_LOZ_N | 4196.125403 | 0.000000e+00 |
| 0 | SME | 4000.017592 | 0.000000e+00 |
| 22 | diff_LOZ_H | 3984.335388 | 0.000000e+00 |
| 13 | abs_delta_LOZ_N | 3227.223007 | 0.000000e+00 |
| 23 | diff_LOZ_F | 2739.500067 | 0.000000e+00 |
| 15 | abs_delta_LOZ_Z | 2704.288296 | 0.000000e+00 |
| 20 | diff_LOZ_E | 2637.581805 | 0.000000e+00 |
| 21 | diff_LOZ_Z | 2637.421331 | 0.000000e+00 |
| 17 | diff_delta_LOZ_E | 2633.139115 | 0.000000e+00 |
| 14 | abs_delta_LOZ_E | 2632.865007 | 0.000000e+00 |
| 18 | diff_delta_LOZ_Z | 2631.087540 | 0.000000e+00 |
| 1 | SMR | 1971.934935 | 0.000000e+00 |
| 10 | delta_LOZ_N | 1920.889711 | 0.000000e+00 |
| 7 | LOZ_F | 1762.486811 | 0.000000e+00 |
| 11 | delta_LOZ_E | 1561.601176 | 4.137389e-316 |
| 5 | LOZ_Z | 1347.979149 | 6.599958e-276 |
| 6 | LOZ_H | 1104.974678 | 3.472105e-229 |
| 3 | LOZ_N | 661.679008 | 4.492896e-141 |
| 12 | delta_LOZ_Z | 434.354443 | 2.431597e-94 |
| 9 | LOZ_I | 335.317142 | 1.259375e-73 |
| 4 | LOZ_E | 66.733386 | 3.509178e-16 |
| 8 | LOZ_D | 0.268771 | 6.041698e-01 |

Рисунок 3 – Оценка результатов исследования с помощью метода ANOVA

Результаты анализа показали, что наиболее важным критерием в выборке является параметр AP. Значимость остальных 23 критериев была определена в порядке убывания.

3. Метод взаимной информации — это эффективный инструмент для

изучения взаимосвязи между переменными. Этот метод является основой информационной теории и находит широкое применение во многих областях, включая машинное обучение, статистический анализ и обработку данных [6].

В контексте машинного обучения и анализа данных, метод взаимной информации часто используется для определения важности признаков. Подсчет взаимной информации между признаками и целевой переменной помогает определить, какие признаки наиболее важны для предсказания значения целевой переменной, а какие признаки имеют меньшую значимость и могут быть удалены из анализа для упрощения модели [7].

В процессе нашего исследования мы воспользовались методом взаимной информации, чтобы провести анализ значимости признаков в нашем наборе данных. Этот метод оказался особенно полезным при определении, какие из наших признаков наиболее важны для нашей целевой переменной.

Применив метод взаимной информации, мы были в состоянии вычислить степень связи между каждым из наших признаков и целевой переменной. Это представляло собой процесс создания новой структуры данных, которая содержала все наши признаки вместе с их соответствующими значениями взаимной информации.

| | Feature | Mutual Information |
|---|---|---|
| 23 | diff_LOZ_F | 0.323326 |
| 18 | diff_delta_LOZ_Z | 0.315144 |
| 19 | diff_LOZ_N | 0.314011 |
| 21 | diff_LOZ_Z | 0.311060 |
| 16 | diff_delta_LOZ_N | 0.310196 |
| 22 | diff_LOZ_H | 0.308419 |
| 13 | abs_delta_LOZ_N | 0.288992 |
| 20 | diff_LOZ_E | 0.274832 |
| 17 | diff_delta_LOZ_E | 0.273660 |
| 10 | delta_LOZ_N | 0.263962 |
| 0 | SME | 0.235968 |
| 2 | AP | 0.234868 |
| 14 | abs_delta_LOZ_E | 0.222166 |
| 15 | abs_delta_LOZ_Z | 0.213449 |
| 11 | delta_LOZ_E | 0.185400 |
| 12 | delta_LOZ_Z | 0.179438 |
| 1 | SMR | 0.153017 |
| 7 | LOZ_F | 0.150333 |
| 5 | LOZ_Z | 0.140962 |
| 6 | LOZ_H | 0.137886 |
| 3 | LOZ_N | 0.123658 |
| 8 | LOZ_D | 0.120479 |
| 9 | LOZ_I | 0.106063 |
| 4 | LOZ_E | 0.103619 |

Рисунок 4 – Результаты на основе метода взаимной информации

На основе результатов можно сделать следующие выводы:

Признаки "diff_LOZ_F", "diff_delta_LOZ_Z", "diff_LOZ_N", "diff_LOZ_Z", "diff_delta_LOZ_N" и "diff_LOZ_H" имеют наибольшую взаимную информацию с целевой переменной "Zenith". Это означает, что эти признаки, возможно, наиболее важны для предсказания значения "Zenith".

Мы проанализировали полученные результаты, что позволило нам сделать выводы о важности каждого признака. Этот анализ помог определить, какие признаки, возможно, являются наиболее важными для предсказания нашей целевой переменной и какие признаки, вероятно, имеют меньшую важность и могут быть исключены для упрощения нашей модели.

В ходе использования вышеописанных методов нами было выявлено, что такие признаки как “diff_LOZ_N”, “diff_LOZ_F” и “AP” являются самыми важными в задаче бинарной классификации для прогнозирования полярных сияний.

Заключение

В процессе нашего исследования мы изучили множество методов, используемых для определения значимости признаков в контексте бинарной классификации. Оценка полученных данных подтвердила, что эти подходы представляют собой согласованные инструменты для определения важности признаков.

**Библиографический список:**

1. А. В. Воробьев, В. А. Пилипенко, Т. А. Еникеев, Г. Р. Воробьева, О.И. Христодуло. Система динамической визуализации геомагнитных возмущений по данным наземных магнитных станций (2021). Научная визуализация 13.1: 162 - 176, DOI: 10.26583/sv.13.1.11.

2. Воробьев А.В., Пилипенко В. А.. Подход к восстановлению геомагнитных данных на базе концепции цифровых двойников. Солнечно-земная физика. 2021. Т. 7, No 2. С. 53–62. DOI: 10.12737/szf- 72202105.

3. Воробьев А.В., Пилипенко В.А., Сахаров Я.А., Селиванов В.Н. Статистические взаимосвязи вариаций геомагнитного поля, аврорального электроджета и геоиндуцированных токов. Солнечно-земная физика. 2019. Т. 5, No 1. С. 48–58. DOI: 10.12737/szf-51201905.

4. Vorobev, A. V., V. A. Pilipenko, R. I. Krasnoperov, G. R. Vorobeva, and D. A. Lorentzen (2020), Short-term forecast of the auroral oval position on the basis of the “virtual globe” technology, Russ. J. Earth. Sci., 20, ES6001, doi:10.2205/2020ES000721.

5. Воробьев, А.В. Геоинформационная система для анализа динамики экстремальных геомагнитных возмущений по данным наблюдений наземных станций / А.В. Воробьев, В.А. Пилипенко, Т.А. Еникеев, Г.Р. Воробьева // Компьютерная оптика. – 2020. – Т. 44, No 5. – С. 782-790. – DOI: 10.18287/2412-6179-CO-707.

6. Коваленко, О. А. (2014). Использование метода взаимной информации для анализа зависимостей в данных. Вестник Компьютерных и

Информационных Технологий, 3(35), 30-34.

7. Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review E, 69(6), 066138.

8. Миллер, Р. Г. (2007). Наглядная статистика: использование Microsoft Excel. Питер.

9. Montgomery, D. C. (2017). Design and analysis of experiments. John Wiley & Sons.

10. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

11. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

12. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

13. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.