Анализ и прогнозирование транспортировки пассажиров в альтернативное измерение

Дарья Попова 1

*1- Технический университет Молдовы (UTM)*

g-mail: daria.popova392@gmail.com

**Абстракт**

В данной статье представлен анализ данных на основе личных записей, извлеченных из поврежденной компьютерной системы корабля. Для более полного понимания данных и выявления ключевых трендов был проведен анализ данных исследовательским методом (EDA), включая графическое представление, были выявлены потенциальные паттерны и зависимости между различными переменными. Затем построение моделей машинного обучения для предсказания вероятности транспортировки.

**Введение**

Данное исследование направлено на повышение вероятности предсказания транспортировки пассажиров в альтернативное измерение во время столкновения космического корабля "Титаник" с пространственно-временной аномалией. Прогнозирование транспортировки и выявление главных влияющих факторов поможет спасти сотни жизней.

Целью проекта является комплексный анализ распределения пассажиров, идентификация факторов, влияющих на транспортировку, а также разработка моделей для предсказаний вероятностей. Было представлено 8630 личных записей, извлеченных из поврежденной компьютерной системы корабля, что предоставляет уникальную возможность получить глубокое понимание событий, произошедших в ходе столкновения космического корабля "Титаник" с пространственно-временной аномалией. При таком масштабе исследования можно выделить не только общие тенденции, но и уникальные особенности, связанные с инцидентом.

Для достижения этих целей были использованы различные методы статистического анализа и машинного обучения, включая логистическую регрессию, метода случайного леса и метода опорных векторов. Эти методы помогли исследовать взаимосвязи между различными аспектами транспортировки, такими как возраст, планета, с которой пассажир вылетел, сон, номер каюты и их значимость в ходе эксперимента.

Таким образом, данное исследование представляет собой важный шаг в направлении более эффективного прогнозирования вероятности транспортировки, что несомненно способствует уменьшению потенциального ущерба и спасению жизней.

**Материалы и Методы**

Для исследования был использован набор личных записей, извлеченных из поврежденной компьютерной системы корабля. Они включают информацию о самих пассажирах, их возраст, статус, планеты, из который прибыли, а также другие важные личные аспекты.

В начале исследования была проведена предварительная очистка данных для устранения любых аномалий или неполных записей. Затем, для анализа взаимосвязей между различными признаками пассажиров, были использованы методы корреляционного анализа и визуализации данных. Все операции проводились при помощи языка R.

Следующим шагом было создание и обучение моделей для прогнозирования вероятности транспортировки. Здесь были применены различные методы машинного обучения, включая логистическую регрессию, метода случайного леса и метода опорных векторов. Весь датасет был разделён на обучающую (70%) и тестовую (30%) выборки.

Важной частью проекта стало и тестирование выбранных моделей. Для этого использовались различные методы анализа ошибок, оценки точности и надежности предсказаний моделей: изучение признаков, поcтроение матрицы ошибок, а также метрики оценки — точность (Accuracy), полнота (Recall), Precission, F1-Score.

**Результаты**

В ходе проекта были получены результаты, которые могут быть использованы для дальнейших исследований данной тематики. В выбранном датасете 13 признаков, из которых 6 числовых и 7 категориальных. Ниже будет представлено их подробное описание:

* PassengerId - уникальный идентификатор для каждого пассажира.
* HomePlanet - планета, с которой пассажир вылетел, обычно это планета его постоянного проживания.
* CryoSleep - Указывает, желает ли пассажир быть переведенным в состояние условной анимации на время путешествия. Пассажиры, находящиеся в криосне, находятся в своих каютах.
* Cabin - Номер каюты, в которой находится пассажир
* Destination - планета, на которой пассажир будет высаживаться.
* Age - Возраст пассажира.
* VIP - Оплатил ли пассажир специальное VIP-обслуживание во время рейса.
* RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Сумма, которую пассажир оплатил в каждом из многочисленных роскошных сервисов космического корабля "Титаник".
* Name - Имя и фамилия пассажира.
* Transported - Был ли пассажир перемещен в другое измерение. Это цель, столбец, который вы пытаетесь предсказать.

Первым делом необходимо было представить данные типа «character» в «factor», так как многие статистические модели в R работают более эффективно с факторными переменными. Также проверено количество пропущенных значений. В наборе данных пропущено около 2% значений, это не сильно повлияет на результаты нашего анализа, но данные было решено не удалять. Численные признаки были заменены средним значением, а категориальные – модой. Также зависимая переменная «Transported» распределена равномерно, см.рис.1. Следующими шагами было анализ влияния других переменных на зависимую.

Изображение выглядит как текст, снимок экрана, Шрифт, диаграмма

Автоматически созданное описание

**Рис. 1 Распределение зависимой переменной**

Переменная «Name» не несла никакой информации, поэтому её пришлось удалить. Признак «Cabin», имеющий вид deck/num/side, где side может быть либо P – Port(левый борт), либо S – Starboard (капитанский борт правый), был разделён на 3 части: Deck, Deck\_Number и Side. После дальнейшего анализа было замечено, что Deck\_Number не несёт никакой информативности, с ним пришлось распрощаться. С другой стороны, переменная Deck показала, что больше всего заселено в части F, но вероятность быть транспортированным у группы c меньшим количеством людей B, это можно увидеть на рисунке 2.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

**Рис. 2 Распределение места проживания**

Также было замечено, что признак «CryoSleep» очень важен, так как шанс того, что пассажир будет перемещён, куда выше, если он находится в криосне. VIP статус никак не повлияет будет ли человек за бортом первым в очереди на транспортировку, привилегий не существует. Также благодаря визуализации на гистограмме была замечена самая частая возрастная категория – это от 20 до 25 (рис. 3).

Изображение выглядит как График, диаграмма, линия

Автоматически созданное описание

**Рис. 3 Гистограмма возраста**

Признак «Passenger\_id» в неподходящем формате, где каждый идентификатор имеет вид gggg\_pp, где gggg означает группу, с которой путешествует пассажир, а pp - его номер в группе. Было решено срезать часть его номера группы и оставить только число, характеризующее его группу на борту. После выяснилось, что максимальная группа состоит из 8 человек, больше всего пассажиров находятся по одиночке, но выше шанс улететь, когда пассажир находится в третьей группе.

После анализа зависимостей категориальных признаков с целевой переменной, были проанализированы численные. Построив матрицу корреляции, было замечено, что корреляций нет, ни положительных, ни отрицательных, где показано на рис.4, это означает, что нет чёткой тенденции или закономерности.

**Изображение выглядит как текст, снимок экрана, число, диаграмма

Автоматически созданное описание**

**Рис. 4 Матрица корреляции для численных переменных**

Следующим шагом были проанализированы зависимости одного признака на другой, где можно получить общую картину. Так на изображении представлена диаграмма Санки, которая иллюстрирует потоки некоторых величин между важными признаками, планету, из которой приехал пассажир и планету, где он будет высаживаться. В данном случае потоки идут между такими небесными телами, как Марс, Европа, Земля, и TRAPPIST-1e, 55 Cancri e и PSO J318.5-22. Ширина лент пропорциональна количеству представляемого потока. Большее количество землян делает свою остановку в TRAPPIST-1e на рисунке 5.

Изображение выглядит как текст, снимок экрана, линия, дизайн

Автоматически созданное описание

**Рис. 5 График потоков между признаками «HomePlanet» и «Destination»**

Следующим этапом над проектом было построение нескольких моделей предсказания вероятности транспортировки по их важным характеристикам, попробовав различные ассоциации признаков, были выявлены результаты, что модель показывает самые высокие значения, используя признаки:

* HomePlanet - планета, с которой пассажир вылетел, обычно это планета его постоянного проживания.
* CryoSleep - Указывает, желает ли пассажир быть переведенным в состояние условной анимации на время путешествия. Пассажиры, находящиеся в криосне, находятся в своих каютах.
* Deck- Номер каюты, в которой находится пассажир
* Side – Сторона корабля, где находится пассажир
* Destination - планета, на которой пассажир будет высаживаться.
* Age - Возраст пассажира.
* VIP - Оплатил ли пассажир специальное VIP-обслуживание во время рейса.
* RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Сумма, которую пассажир оплатил в каждом из многочисленных роскошных сервисов космического корабля "Титаник".

Эти переменные были выбраны неслучайно, так как они показали наибольшее влияние на графиках визуализации.

Данные были разделены на обучающую и тестовую модели в соотношении 70% к 30%, где 6084 данных стали df\_train, а 2609 – df\_test. Также категориальные признаки 'Transported', 'CryoSleep' и 'VIP', представленные в виде TRUE/FALSE были преобразованы в формат 0/1.

Также была представлена таблица статистических результатов регрессионной модели из языка программирования R, которая отформатирована функцией tidy() из пакета broom. В таблице представлены переменные модели вместе с их оценками, стандартными ошибками, t-статистикой и p-значениями. Значимые переменные, такие как HomePlanetEuropa, CryoSleep и DeckC, имеют низкие p-значения, что указывает на их статистическую значимость в модели. Переменные, такие как DeckD, DeckE и DeckF, не не столь статистически значимы на основе их p-значений. Эти результаты предоставляют важную информацию о факторах, влияющих на исследуемую зависимую переменную.

Изначально была построена модель множественной логистической регрессии. Она показала значения точности (Accuracy) 0.80528, что является очень хорошими показателями для данной задачи. Позже была реализована матрица ошибок (confusion matrix) для модели классификации. В нашей матрице ошибок такие результаты:

True Positive (TP): Модель правильно предсказала положительный класс (True) - в данном случае, 1078 наблюдения.

True Negative (TN): Модель правильно предсказала отрицательный класс (False) - 1016 наблюдений.

False Positive (FP): Модель неправильно предсказала положительный класс – 236 наблюдений.

False Negative (FN): Модель неправильно предсказала отрицательный класс - 251 наблюдение.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

**Рис. 4 Матрица ошибок**

Также были выведены дополнительные метрики для анализа модели, которые предоставили следующие результаты:

* Полнота (Recall): 0.794
* Специфичность: 0.812
* Точность (Precision): 0.82
* F1-мера: 0.807

Модель имеет высокую полноту в 79.4% и специфичность в 81.2%, что свидетельствует о хорошей способности корректно выявлять истинно положительные и истинно отрицательные случаи. Точность составляет 82%, а F1-мера - 80.7%, подчеркивая баланс между точностью и полнотой, что говорит о хорошей общей эффективности модели классификации.

Для улучшения предсказаний была выбраны ещё 2 модели: метода случайного леса и метода опорных векторов.

Их результаты составили Accuracy случайного леса: 0.81

Accuracy метода опорных векторов: 0.807

Эти результаты свидетельствуют о хорошем качестве всех рассмотренных моделей, что делает их перспективными для дальнейших исследований и использования в практических приложениях.

**Обсуждение**

Исследование подчеркивает важность комплексного анализа данных о судьбе пассажиров космического корабля «Титаник», которые были перемещены в параллельные вселенные в результате столкновения с пространственно-временной аномалией. Этот анализ необходим для разработки более точных методов прогнозирования и дальнейшего нахождения пассажиров. Однако, следует учитывать ограничения данных и потенциальные искажения. Дальнейшие исследования могут включать в себя более глубокий анализ взаимосвязи различных параметров и разработку более совершенных моделей для прогнозирования и успешного спасения тех, кто оказался в других измерениях.

**Код**

Код можно найти по ссылке [здесь](https://github.com/Dariushkaa/DS/blob/main/Data/Article.docx)

**Данные**

Данные найти по ссылке [здесь](https://github.com/Dariushkaa/DS/blob/main/Spaceship.csv)