Анализ и прогнозирование магнитуды землетрясений

Мария Смарчевская 1

*1- Технический университет Молдовы*

e-mail: maria.smarcevscaia@iis.utm.md

**Абстракт**

В данной статье представлен анализ данных о землетрясениях на основе обширного набора данных, охватывающего период с 1990 по 2023 год. Был проведён разведочный анализ данных, включая графическое представление, выбор наиболее важных переменных, обнаружение отклонений и аномалий, а затем построение моделей машинного обучения для предсказания значений магнитуды по некоторым характеристикам.

**Введение**

Данное исследование направлено на улучшение предсказания магнитуды землетрясений, что может спасти жизни и предотвратить разрушения. Землетрясения — это одно из самых опасных природных явлений, и их анализ может значительно повысить эффективность мер реагирования на чрезвычайные ситуации, поэтому статья освещает ключевые аспекты, связанные с анализом и предсказанием магнитуды землетрясений по всему миру, охватывая период с 1990 по 2023 год.

Целью проекта является комплексный анализ пространственного и временного распределения землетрясений, идентификация факторов, влияющих на их магнитуду, а также разработка моделей для ее предсказаний. Проект использует обширный набор данных, включающий более 3 миллионов записей о землетрясениях, что позволяет провести детальный анализ и выявить ключевые тенденции.

Для достижения этих целей были использованы различные методы статистического анализа и машинного обучения, включая линейную регрессию, построение дерева решений и случайного леса. Эти методы помогли исследовать взаимосвязи между различными аспектами землетрясений, такими как магнитуда, глубина, географическое положение и их значимость. Особое внимание уделено анализу того, как эти факторы взаимодействуют, чтобы повысить точность прогнозов.

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

**Материалы и Методы**

Для исследования были использованы данные о землетрясениях, предоставленные Геологической службой США (USGS). Эти данные охватывают период с 1990 по 2023 год и включают информацию о магнитуде, глубине, времени и местоположении землетрясений, а также другие важные геологические и географические параметры.

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

Следующим шагом было создание и обучение моделей для прогнозирования характеристик землетрясений. Здесь были применены различные методы машинного обучения, включая линейную регрессию, построение дерева решений и алгоритм случайного леса. Эти модели были обучены на основе исторических данных, чтобы оценить их способность предсказывать магнитуду будущих землетрясений. Для этого были использованы 50% случайно выбранных данных из набора, затем они были разделены на обучающую (70%) и тестовую (30%) выборки.

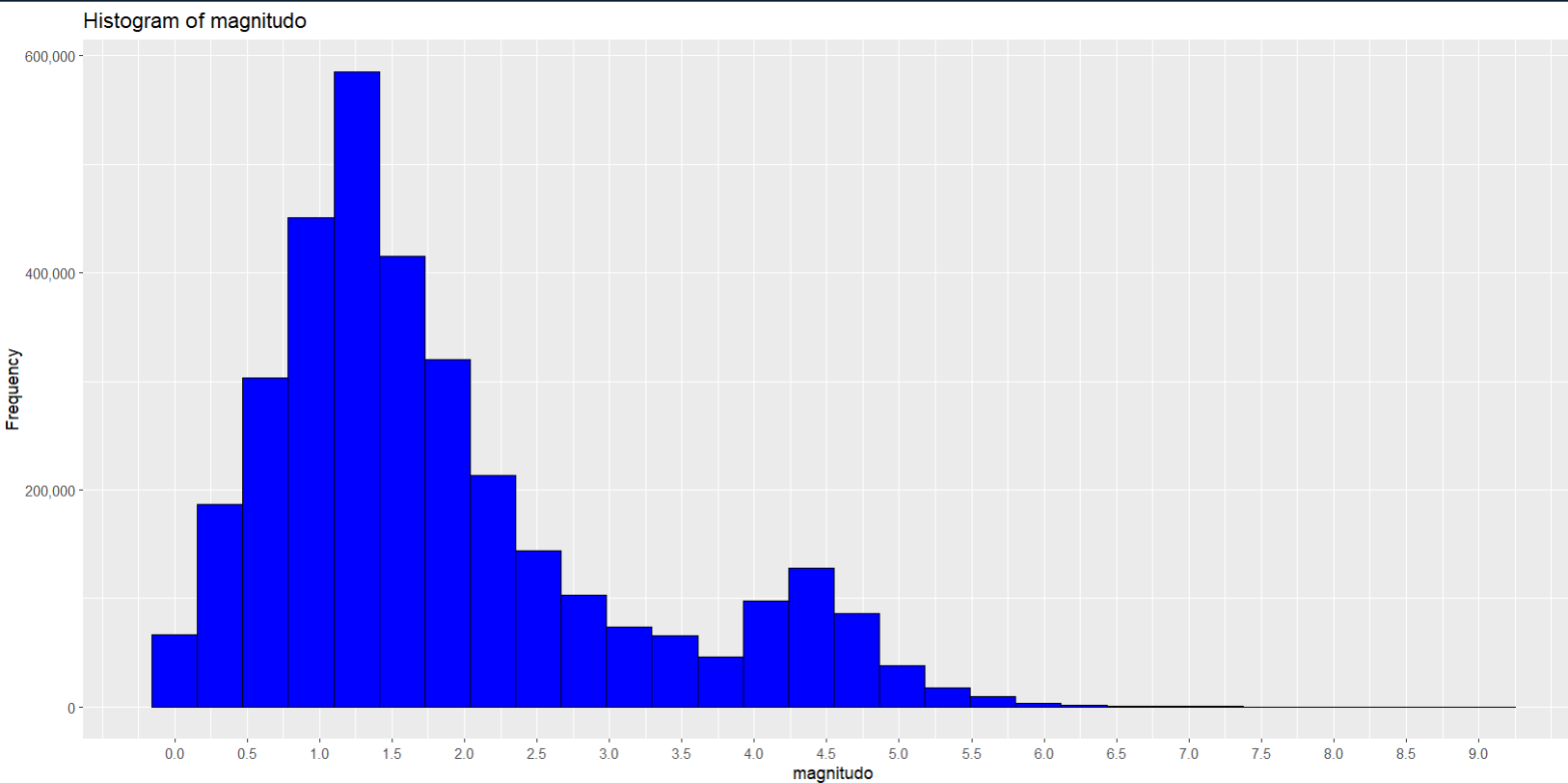
Важной частью проекта стало и тестирование выбранных моделей. Для этого использовались различные методы анализа ошибок, оценки точности и надежности предсказаний моделей: изучение коэффициентов, построение графика остатков и графика предсказаний, а также метрики оценки — среднеквадратическая ошибка (MSE) и корень из среднеквадратической ошибки (RMSE).

**Результаты**

В ходе проекта были получены результаты, описанные в этой части статьи. Данные результаты могут быть использованы для улучшения существующих моделей прогнозирования землетрясений.

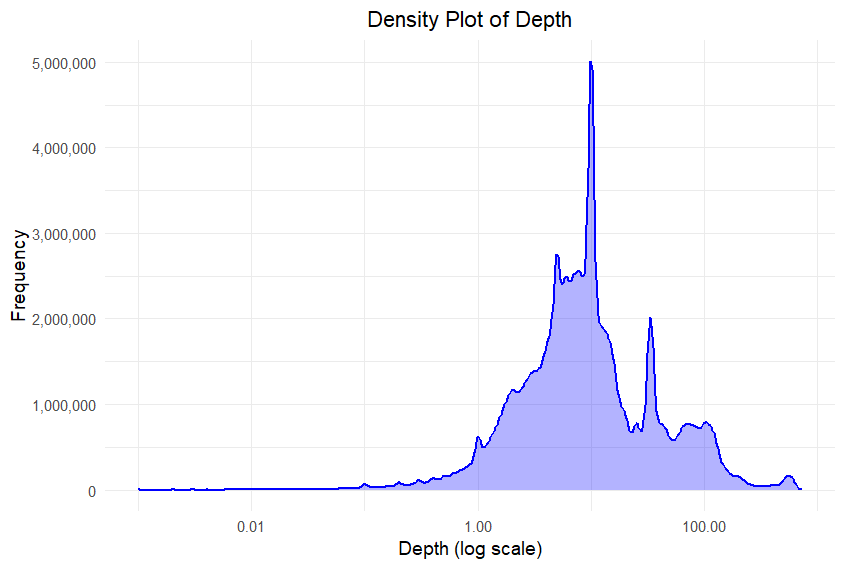
Во время очистки набора данных были обнаружены выбросы в виде отрицательной магнитуды, позже удалённой. Колонка, показывающая, было ли вызвано цунами, не несла никакой информации, поэтому её тоже пришлось убрать. Поля, представляющие различные американские штаты, были заменены на название самого государства, благодаря чему, появилась новая колонка, отображающая названия стран.

Обработка данных показала, что магнитуда и глубина имеют ненормальное распределение, это видно на рис. 3 и рис. 4. Для магнитуды распределение имеет 2 пика, котороые находятся в значениях 1.25 и 4.5 по шкале Рихтера, которые встречаются в примерно 600 000 и 150 000 землетрясениях, соответственно.



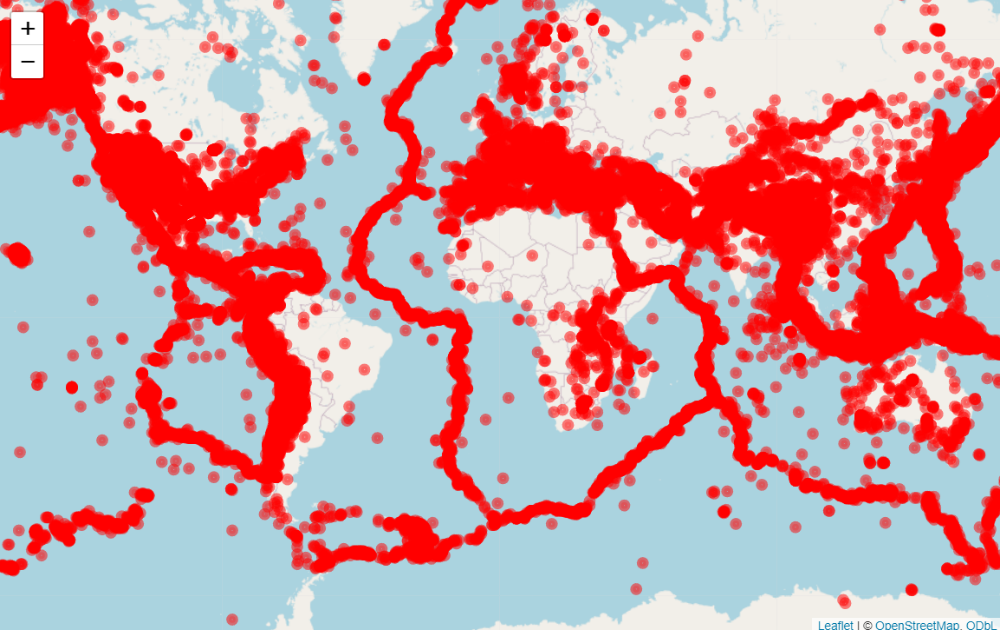
**Рис. 1 Распределение магнитуды**

График распределения для плотности показывает несколько пиков, что может указывать на то, что некоторые глубины встречаются чаще других. Наиболее выраженный пик представляет наиболее распространенную глубину землетрясений в анализируемом наборе данных. Наличие нескольких пиков может свидетельствовать о группировках глубин, что может быть связано с определенными геологическими особенностями или типами сейсмической активности.

****

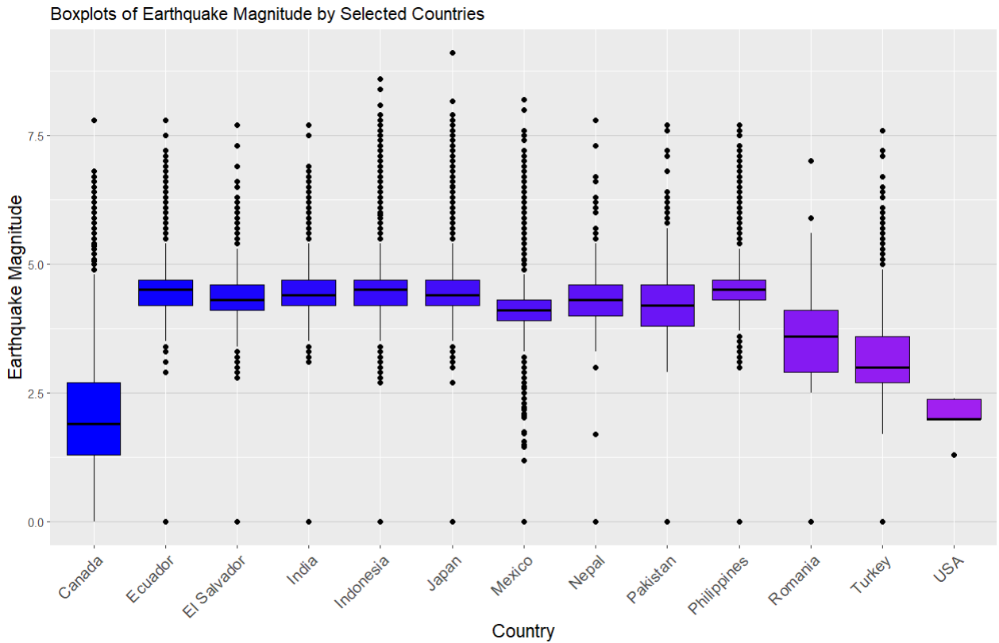
**Рис. 2 График плотности распределения глубины землетрясений в логарифмическом масштабе**

После проведения некоторых исследований было решено, что, возможно, данное распределение у магнитуды и глубины вызвано особенностью набора данных, так как количество строк для каждой страны не совпадает, т.е. данные не сбалансированы территориально.

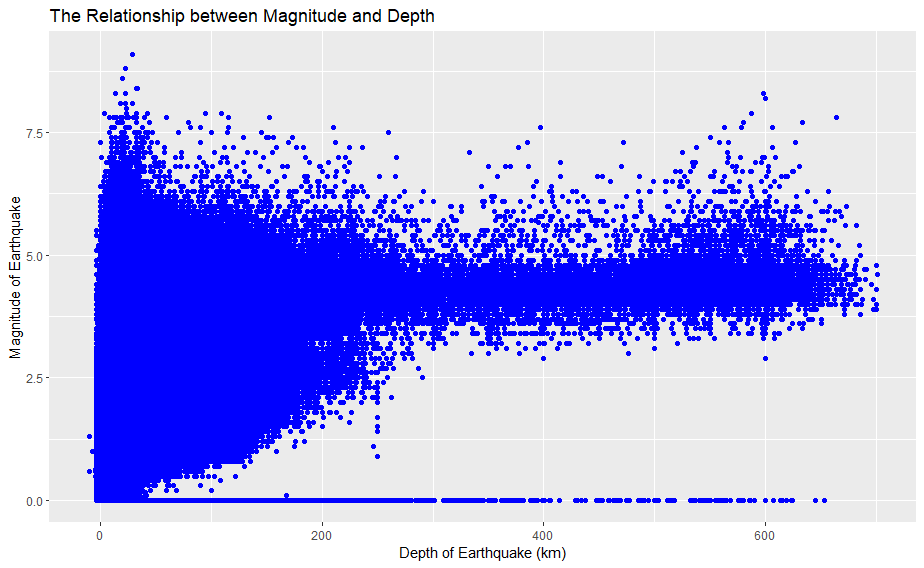
Результаты анализа показали, что большинство значительных землетрясений сосредоточено в определенных географических регионах, благодаря чему получилось построить карту землетрясений на рис. 2, по контурам совпадающую с тектоническими плитами, что подтверждает правильность используемых данных.

**Рис. 2 Карта землетрясений**

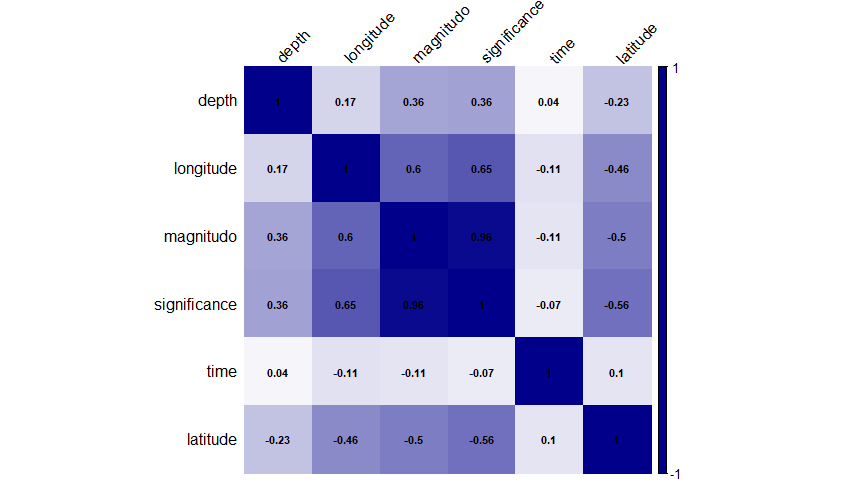
Позже была проанализирована магнитуда по конкретным странами построен соответствующий график на рис.3, на котором можно заметить определённую тенденцию. В странах, где часто происходят землетрясения, среди которых Эквадор, Сальвадор, Индия, Япония и некоторые другие, отмечается большое количество выбросов с магнитудой землетрясений со значением до 9.5 по шкале Рихтера.



**Рис. 3 Распределение магнитуды по выбранным странам**

Было выявлено, что глубина и магнитуда землетрясений не имеют четкой взаимосвязи. Данные разбросаны по всему спектру глубин с большим скоплением землетрясений на меньших глубинах (ближе к 0 км). Магнитуда этих землетрясений варьируется в широком диапазоне от ниже 1.0 до выше 7.5 по шкале Рихтера. На построенном точечном графике на рис. 4 взаимосвязи этих полей заметен узор, где диапазон магнитуд, кажется, шире на меньших глубинах и сужается с увеличением глубины. Однако, землетрясения с большей магнитудой все еще происходят на больших глубинах.

**Рис. 4 Взаимосвязь между глубиной и магнитудой**

Отсутствие чёткой тенденции или закономерности предполагает, что нет сильной линейной корреляции между глубиной и магнитудой землетрясений, поэтому была построена матрица корреляции между магнитудой и другими числовыми характеристиками на рис. 5.

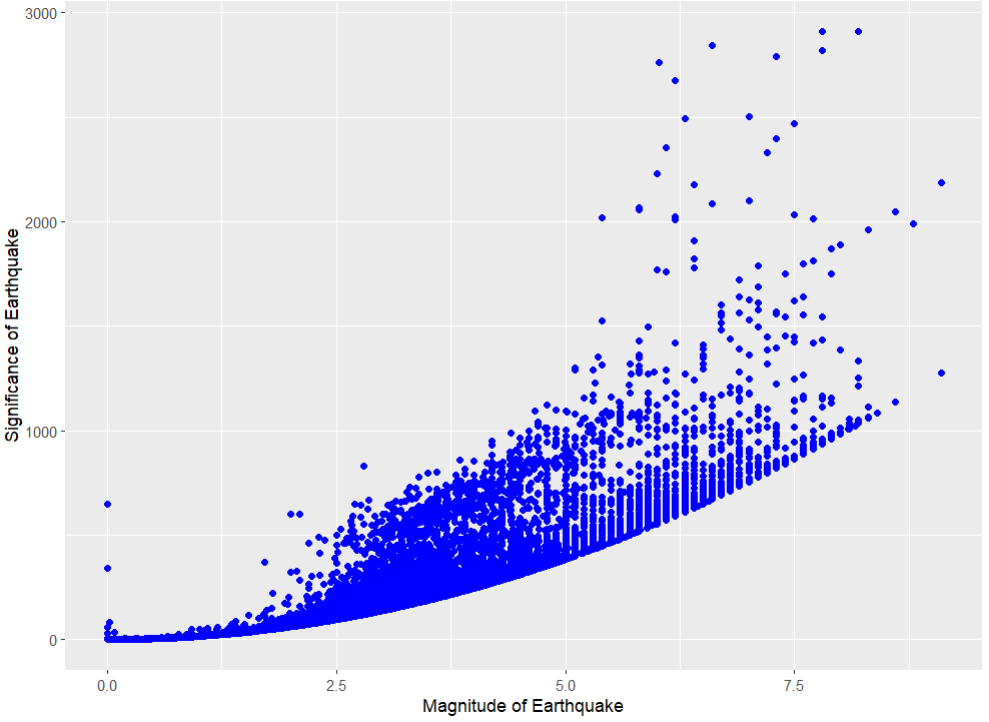
**Рис. 5 Корреляция между магнитудой и другими характеристиками**

Анализ показал следующую корреляцию:

* между магнитудой и глубиной ─ 0.36
* между магнитудой и широтой ─ -0.5
* между магнитудой и долготой ─ 0.6
* между магнитудой и значимостью ─ 0.96
* между магнитудой и длительностью ─ -0.11

Стоит отметить, что очень высокая корреляция между магнитудой и значимостью ожидаема, так как значимость напрямую от нее зависит. Чем выше магнитуда землетрясения, тем большее значение (и, возможно, влияние или последствия) оно имеет. Была построена диаграмма взаимосвязи между ними на рис. 6, где отмечено, что корреляция кажется нелинейной, так как увеличение значимости не происходит равномерно с увеличением магнитуды; вместо этого, значимость растёт более стремительно на более высоких значениях магнитуды.

Точечная диаграмма также показывает, что для небольших землетрясений значимость остаётся относительно низкой и начинает экспоненциально возрастать после определённого порога магнитуды (около 1).



**Рис. 6 Взаимосвязь между значимостью и магнитудой**

В конце работы над проектом получилось построить несколько моделей предсказания магнитуды по их важным характеристикам:

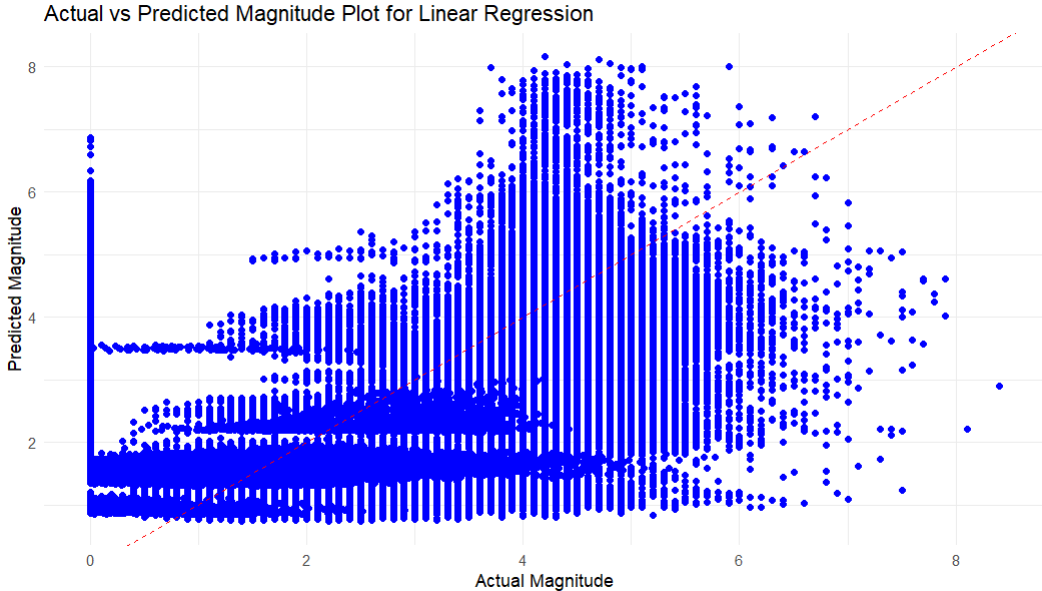
* долгота (longitude)
* широта (latitude)
* глубина (depth)
* длительность (time)

Эти переменные были выбраны неслучайно, так как они показали наибольшее влияние на первую линейную модель и корреляцию с магнитудой.

Изначально была построена модель множественной линейной регрессии. Она показала значения MSЕ ─ 0.81 и RMSE ─ 0.90, что является не очень хорошими показателями для данной задачи, так как отличие предсказанной магнитуды в 0.90 от реальной магнитуды может стоить определённых потерь.

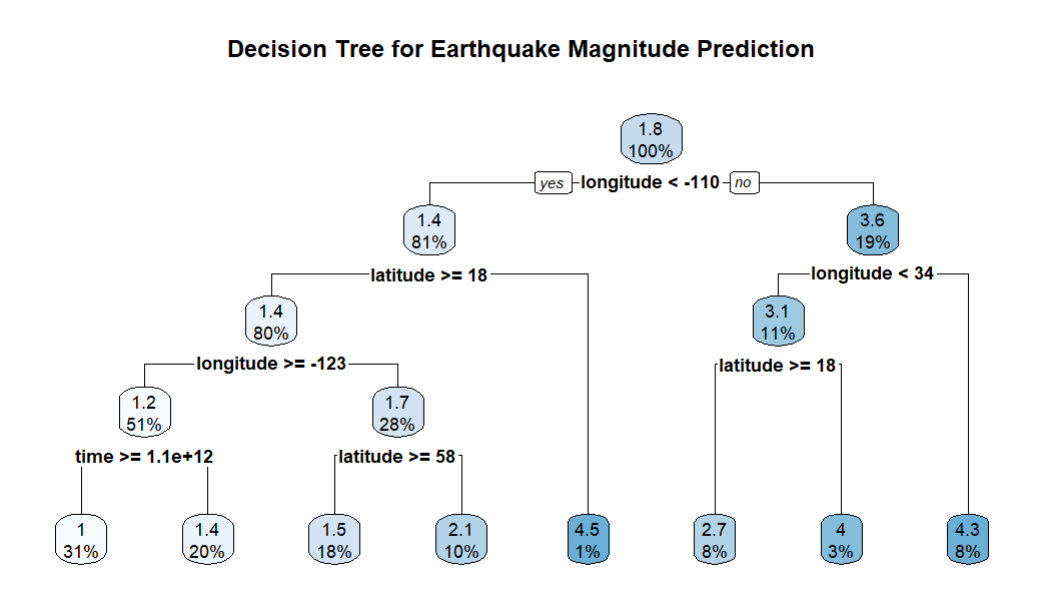
Все предикторы показали большую значимость и p-value < 2e-16. R2 составил 0.48, что означает, что около 48% данных можно объяснить этой моделью.

На графике предсказаний на рис. 7 отсутствует чёткое направление или тренд, который следует ожидать от хорошо работающей регрессионной модели. Существует большое количество предсказаний с одинаковыми или очень близкими значениями, что приводит к вертикальным линиям на графике.



**Рис. 7 Сравнительный график фактической и предсказанной магнитуды для линейной регрессии**

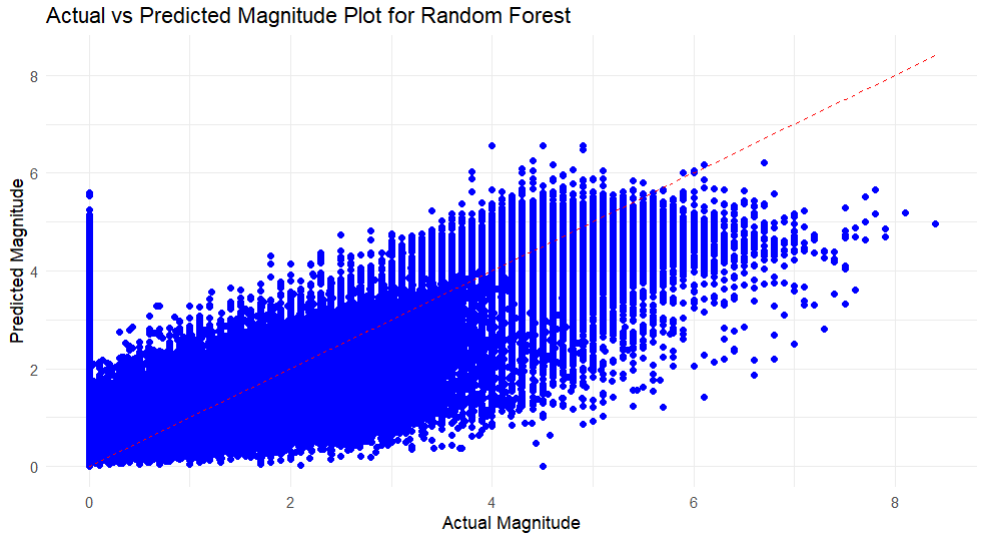
Для улучшения предсказаний была выбрана новая модель – дерево решений на рис. 8. Каждый узел и лист дерева содержит числовые значения, которые, обозначают магнитуду землетрясений и процент случаев, соответствующих каждому условию.

Дерево решений дало лучшие результаты метрик, MSЕ ─ 0.50 и RMSE ─ 0.71.

**Рис. 8 Дерево решений для прогнозирования магнитуды землетрясений**

Не смотря на показатели дерева решений, модель можно улучшить и для этого был построен случайный лес из 5 деревьев решений.

Был построен график предсказанных значений для этой модели на рис. 9. Он показал, что многие точки сосредоточены вдоль линии идеального соответствия, особенно при низких значениях магнитуды, что указывает на то, что модель делает точные предсказания в этом диапазоне. Однако по мере увеличения фактической магнитуды точки распространяются шире, что показывает большую неопределенность или ошибку в предсказаниях при более высоких магнитудах. также модель с ошибками предсказывает нулевую магнитуды.

**Рис. 9 Сравнительный график фактической и предсказанной магнитуды для случайного леса**

Эта модель показала самый лучшие значения метрик: MSE составила ─ 0.30, а RMSE ─ 0.55, и это небольшое отклонение и лучшее, возможно, которого можно добиться на основе используемого датасета.

**Обсуждение**

Исследование подчеркивает значимость комплексного анализа данных о землетрясениях для разработки более точных методов прогнозирования. Однако, необходимо учитывать ограничения данных и потенциальные искажения. Дальнейшие исследования могут включать более глубокий анализ взаимосвязи между различными параметрами и разработку более совершенных моделей для прогнозирования.

**Код и данные**

Код и данные проекта можно найти по ссылке [Whowhywhere/DS: for data science (github.com)](https://github.com/Whowhywhere/DS)