# Аналитический отчёт по курсовой работе

## 1. Введение

Цель проекта – применение классических методов машинного обучения для прогноза активности (IC50), токсичности (CC50) и избирательности (SI) химических соединений против вируса гриппа на основе их дескрипторов.

**IC₅₀, CC₅₀ и SI** — это ключевые параметры, используемые в фармакологии и токсикологии для оценки эффективности и безопасности соединений (например, лекарств, противовирусных агентов).

**1. IC₅₀ (Half-Maximal Inhibitory Concentration)**

* **Что это**: Концентрация вещества, необходимая для **подавления специфической биологической активности на 50%**.
* **Применение**:
  + Используется для оценки **эффективности** препарата. Например:
    - Для противовирусного средства IC₅₀ показывает, какая концентрация требуется, чтобы заблокировать репликацию вируса на 50%.
    - В ферментативных исследованиях — концентрация, подавляющая активность фермента на 50%.
* **Чем ниже IC₅₀**, тем выше эффективность (меньшая концентрация нужна для действия).

**2. CC₅₀ (Half-Maximal Cytotoxic Concentration)**

* **Что это**: Концентрация вещества, вызывающая **гибель 50% клеток** (например, здоровых клеток организма).
* **Применение**:
  + Отражает **токсичность** соединения.
  + **Чем выше CC₅₀**, тем безопаснее препарат (требуется большая доза для повреждения клеток).

**3. SI (Selectivity Index / Индекс селективности)**

* **Что это**: Отношение **CC₅₀ к IC₅₀**, показывающее, насколько препарат селективен в воздействии на мишень (например, вирус) по сравнению с токсичностью для клеток.
* **Интерпретация**:
* SI > 10: Препарат считается безопасным и селективным (эффективен при низких концентрациях, малотоксичен).
* SI < 1: Токсичность превышает эффективность (опасен для применения).

## 2. Разведочный анализ данных (EDA)

• Датасет - 998 соединений, 195 дескрипторов, три целевые переменные (IC50, CC50, SI).  
• Удалены признаки, содержащие только 0, константные признаки, пропуски, а также дубликаты.  
• Применено логарифмическое преобразование распределений.  
• Spearman корреляции: IC50↔CC50 ≈ 0.62, IC50↔SI ≈ -0.64, CC50↔SI ≈ 0.53.  
• Ни один признак не имеет сильной корреляции с целевыми переменными.

3. Регрессия: прогноз IC50

Linear Regression

R² = 0.084:

Это означает, что модель объясняет только 8.4% дисперсии зависимой переменной. Такой низкий R² указывает на то, что линейная модель не способна уловить большую часть закономерностей в данных.

MAE = 262.15:

Средняя абсолютная ошибка равна 262.15 единиц. В среднем модель ошибается на ≈262.15 единиц при прогнозировании.

MSE = 181735.92:

Среднеквадратичная ошибка относительно велика, что говорит о наличии существенных ошибок прогнозирования, особенно если в данных присутствуют выбросы (так как большие ошибки сильно влияют на MSE).

Вывод: Линейная регрессия показывает слабое качество предсказания, что может быть связано либо с недостаточной линейностью зависимости между признаками и целевой переменной, либо с тем, что важные факторы не включены в модель.

Random Forest

R² = 0.410:

Модель объясняет около 41% дисперсии целевой переменной. Это значительное улучшение по сравнению с линейной регрессией, что говорит о лучшей способности модели улавливать сложные зависимости.

MAE = 205.98:

Средняя абсолютная ошибка снизилась до 205.98 единиц, что свидетельствует о более точных прогнозах.

MSE = 116945.44:

Меньшее значение MSE по сравнению с линейной регрессией указывает на снижение серьезных ошибок.

Вывод: Модель случайного леса способна лучше моделировать закономерности данных, вероятно, за счет использования ансамбля деревьев, который учитывает нелинейные зависимости и взаимодействия между признаками.

XGBoost

R² = 0.278:

Модель объясняет около 27.8% дисперсии зависимой переменной, что выше, чем у линейной регрессии, но ниже, чем у случайного леса.

MAE = 214.46:

Средняя абсолютная ошибка немного больше, чем у случайного леса, что говорит о менее точных прогнозах.

MSE = 143219.01:

Значение MSE также выше, чем в случае случайного леса, что может указывать на наличие нескольких больших ошибок.

Вывод: XGBoost проявляет себя лучше, чем простая линейная регрессия, но в данном случае его показатели уступают случайному лесу. Возможно, параметры модели или особенности данных требуют дополнительного тюнинга.

LightGBM

R² = 0.382:

Объясняется около 38.2% дисперсии зависимой переменной, что сопоставимо с результатами XGBoost и немного ниже случайного леса.

MAE = 204.79:

Средняя абсолютная ошибка чуть меньше, чем у случайного леса, что свидетельствует о хорошей точности предсказаний на "средних" ошибках.

MSE = 122551.46:

Значение MSE немного выше, чем у случайного леса, что может говорить о наличии некоторых выбросов или больших ошибок, влияющих на этот показатель.

Вывод: Модель LightGBM показывает сравнимую с Random Forest способность предсказания. Она демонстрирует схожие метрики точности — немного уступая по R², но показывая чуть лучшие MAE. Это указывает на хорошее качество модели, особенно в плане средней ошибки.

Общая интерпретация:

Сравнение моделей по R²:

Random Forest (R²=0.410) показывает наилучший результат с точки зрения объяснения дисперсии целевой переменной, что означает, что данная модель лучше учитывает взаимосвязи между признаками.

LightGBM (R²=0.382) находится немного позади Random Forest, а XGBoost (R²=0.278) уступает обеим ансамблевым моделям.

Линейная регрессия (R²=0.084) дает крайне низкое значение, что свидетельствует о слабой аппроксимации зависимостей.

Ошибка предсказания (MAE и MSE):

Меньшие значения MAE и MSE для Random Forest, XGBoost и LightGBM относительно линейной регрессии указывают на снижение средней абсолютной и квадратичной ошибки, что свидетельствует о более точном прогнозировании.

Особое внимание стоит уделить тому, что MSE чувствителен к большим ошибкам (из-за квадратичного возрастания вклада выбросов).

Случайный лес демонстрирует наиболее низкие значения MSE, что доказывает его устойчивость к большим ошибкам.

Общий вывод:

Ансамблевые методы (случайный лес, XGBoost, LightGBM) показывают лучшие результаты по сравнению с линейной регрессией. Это говорит о том, что данные, скорее всего, имеют нелинейные или более сложные взаимосвязи, которые проще уловить моделям, объединяющим результаты нескольких базовых алгоритмов.

При этом, случайный лес в данном случае выглядит лучшим выбором с точки зрения объяснения дисперсии и минимизации крупных ошибок, хотя разница между ним и LightGBM незначительная.

4. Регрессия: прогноз CC50

Linear Regression

R² = 0.246

Модель объясняет около 24.6% разброса целевой переменной. Это довольно низкое значение, которое указывает, что линейная зависимость не способна полноценно уловить закономерности в данных.

MAE = 396.47

Средняя абсолютная ошибка показывает, что в среднем предсказания отклоняются от истинных значений примерно на 396.47 единиц.

MSE = 305790.11

Высокое значение MSE говорит о большом влиянии возможных выбросов или сильно ошибочных предсказаниях, поскольку ошибки возводятся в квадрат.

Вывод: Линейная регрессия в данном случае выступает как базовая модель и показывает слабое качество предсказаний, что может свидетельствовать о наличии сложных (например, нелинейных) взаимосвязей между признаками и целевой переменной.

Random Forest

R² = 0.544

Модель объясняет 54.4% дисперсии целевой переменной, что значительно лучше, чем у линейной регрессии. Это говорит о том, что ансамблевый метод (за счёт объединения множества деревьев) способен уловить больше информации из данных.

MAE = 292.12

Средняя абсолютная ошибка снизилась до 292.12, что говорит о более точных предсказаниях по сравнению с линейной моделью.

MSE = 184965.80

Значительное снижение MSE по сравнению с линейной регрессией указывает на уменьшение крупных ошибок.

Вывод: Случайный лес демонстрирует лучшую способность моделировать сложные зависимости, чем простая линейная регрессия, что приводит к более высоким значениям R² и меньшим ошибкам.

XGBoost

R² = 0.596

Модель объясняет около 59.6% дисперсии целевой переменной — это наилучший показатель из представленных.

MAE = 280.36

Средняя абсолютная ошибка ещё ниже, что свидетельствует о большей точности предсказаний.

MSE = 163709.70

Меньшее значение MSE говорит о снижении влияния крупных ошибок (выбросов) на итоговую ошибку модели.

Вывод: XGBoost показывает наилучшие результаты среди рассмотренных моделей. Высокий R², а также минимальные показатели ошибок указывают на его способность эффективно выявлять сложные нелинейные зависимости и взаимодействия между признаками.

LightGBM

R² = 0.561

Модель объясняет около 56.1% дисперсии целевой переменной, что также существенно выше линейной регрессии и немного уступает XGBoost.

MAE = 278.75

Средняя абсолютная ошибка почти идентична значению XGBoost, что говорит о хорошей точности на уровне средних ошибок.

MSE = 178177.66

Значение MSE немного выше, чем у XGBoost, что может указывать на наличие нескольких более крупных ошибок, хотя в целом результаты остаются приемлемыми.

Вывод: LightGBM тоже демонстрирует хорошие показатели и явно лучше линейной регрессии. Разница между LightGBM и XGBoost небольшая: при чуть лучшем MAE у LightGBM, XGBoost показывает наилучшее значение R² и немного меньший MSE.

Общая интерпретация

Улучшение по сравнению с линейной регрессией:

Ensemble-модели (Random Forest, XGBoost и LightGBM) значительно превосходят линейную регрессию по всем метрикам. Это говорит о том, что данные, скорее всего, содержат нелинейные зависимости и взаимодействия между признаками, которые простая линейная модель не может адекватно описать.

Сравнение ансамблевых методов:

XGBoost показывает наилучшие результаты (наибольшее значение R² и минимальные показатели ошибок), что свидетельствует о его способности наиболее эффективно улавливать сложные зависимости.

LightGBM демонстрирует схожее качество предсказаний, с чуть меньшим R², но немного лучшим MAE.

Random Forest также улучшает показатели по сравнению с линейной регрессией, хотя результаты немного уступают XGBoost и LightGBM.

Заключение:

В рассматриваемой задаче ансамблевые модели значительно превосходят линейную регрессию, что подчеркивает важность выбора модели, способной учитывать нелинейные взаимосвязи в данных. XGBoost выглядит наиболее предпочтительным, но разница между ансамблевыми методами не столь велика, и окончательный выбор может зависеть от специфики задачи, интерпретируемости модели или требований к скорости обучения и прогнозирования.

4.Регрессия: прогноз SI

Linear Regression

R² = -10.217:  
 Очень отрицательное значение говорит о том, что модель предсказывает настолько плохо, что ошибка предсказаний значительно превышает вариативность целевого признака (то есть модель работает хуже, чем простое предсказание среднего значения).

MAE = 231.89:  
В среднем предсказания отклоняются от истинного значения на 231.89 единиц, что достаточно высоко.

MSE = 222394.58:  
Сильное возрастание ошибки при возведении отклонений в квадрат подтверждает наличие крупных ошибок в предсказаниях.

Вывод: Линейная регрессия явно не справляется с задачей: вероятно, зависимость между признаками и целевой переменной является сложной или нелинейной, а простой линейный подход не способен уловить эту структуру.

Random Forest:

R² = -5.582:  
Несмотря на отрицательное значение, оно менее отрицательно, чем у линейной регрессии, что указывает на некоторое улучшение. Однако по-прежнему модель работает ниже базового уровня (прогноз среднего значения).

MAE = 82.82:  
Средняя ошибка уменьшилась до 82.82 единиц, что является значительным улучшением по сравнению с линейной регрессией.

MSE = 130490.29:  
Также снижение MSE означает, что модель совокупно уменьшила ошибки, хотя остаются отдельные крупные ошибки.

Вывод: Метод случайного леса лучше справляется с задачей по сравнению с линейной регрессией, снижая как среднюю абсолютную ошибку, так и квадратичную ошибку. Однако отрицательный R² говорит о том, что даже этот подход не может объяснить достаточную часть дисперсии целевой переменной – предсказания все еще хуже, чем простое использование среднего значения.

XGBoost:

R² = -7.542:  
Отрицательное значение R² вновь указывает на то, что модель не справляется с задачей настолько, чтобы превзойти базовый прогноз среднего значения, хотя по размеру ошибки немного лучше, чем линейная регрессия, но хуже, чем у случайного леса.

MAE = 99.40:  
Средняя абсолютная ошибка немного выше, чем у Random Forest, что свидетельствует о чуть менее точных предсказаниях на среднем уровне.

MSE = 169362.99:  
Более высокий показатель MSE предполагает, что предсказания содержат более крупные отклонения (выбросы или отдельные сильно ошибочные предсказания).

Вывод: Модель XGBoost улучшает качество предсказаний по сравнению с простой линейной моделью, однако её ошибки остаются достаточно высокими, а отрицательный R² подчеркивает, что выбранная конфигурация модели не способна адекватно уловить целевую зависимость.

LightGBM

R² = -7.570:  
Значение близко к XGBoost, что говорит о схожей производительности по объяснению дисперсии в данных. Отрицательный показатель означает, что модель работает хуже простого базового прогноза.

MAE = 118.45:  
Средняя абсолютная ошибка несколько выше, чем у XGBoost и Random Forest, что может свидетельствовать о менее точных средних предсказаниях.

MSE = 169916.77:  
Показатель квадратичной ошибки почти аналогичен XGBoost – остаются отдельные более крупные ошибки.

Вывод: LightGBM демонстрирует результаты, сопоставимые с XGBoost. Хотя она снижает абсолютные и квадратичные ошибки по сравнению с линейной регрессией, отрицательный коэффициент детерминации остаётся индикатором того, что модель не умеет уловить основную структуру данных.

Общая интерпретация результатов

Отрицательные значения R² для всех моделей:  
Это указывает на то, что ни одна из моделей не справилась с задачей так, чтобы объяснить дисперсию целевой переменной лучше, чем базовый метод, который всегда предсказывает среднее значение.

Возможные причины:

Сложные, возможно нелинейные зависимости, которые даже ансамблевые методы не смогли уловить, возможно, из-за неподходящего выбора гиперпараметров или проблемы в разбиении данных.

Сравнение моделей по MAE и MSE:

Random Forest показывает заметное снижение ошибки (MAE ≈ 82.82), что говорит о том, что модель предсказывает ближе к истинным значениям по абсолютной величине.

XGBoost и LightGBM имеют MAE чуть выше, а также более высокий MSE, что указывает на присутствие нескольких случаев с высокими ошибками предсказаний.

Что можно сделать:

Пересмотреть процесс отбора признаков: возможно, добавить новые или устранить нерелевантные.

Заключение:

Все четыре модели демонстрируют недостаточную способность предсказывать целевую переменную, поскольку их ошибки выше, чем у базового прогноза среднего значения (что отражается отрицательным R²). Из ансамблевых методов Random Forest показывает наименьшую абсолютную ошибку (MAE), но даже она не достигает положительного коэффициента детерминации. Это сигнализирует о том, что необходимо более глубокое изучение данных, корректировка предобработки и, возможно, поиск других признаков или моделей, способных лучше уловить структуру исходных данных.

5. Классификация: IC50 > медианы

Logistic Regression

Accuracy = 0.727: Модель правильно классифицирует 72,7% объектов.

Precision = 0.724: Из предсказанных положительных случаев около 72,4% действительно положительные.

Recall = 0.732: Модель находит примерно 73,2% всех реальных положительных примеров.

F1 = 0.728: Сбалансированное значение, подтверждающее умеренное качество модели по обоим аспектам — точности и полноте.

AUC = 0.806: Довольно высокий показатель дискриминационной способности модели, т.е. она хорошо различает классы при изменении порога.

Вывод: Logistic Regression демонстрирует стабильные и сбалансированные показатели, показывая, что при данном пороге модель работает на уровне около 73% точности, а AUC свидетельствует о хорошем различении классов.

Random Forest

Accuracy = 0.727: Тот же уровень общей точности, что и у логистической регрессии (72,7%).

Precision = 0.729: Чуть выше, чем у Logistic Regression — около 72,9% предсказанных положительных действительно положительны.

Recall = 0.722: Чуть ниже (72,2%), что означает незначительное снижение способности обнаруживать все положительные случаи по сравнению с Logistic Regression.

F1 = 0.725: Схожее значение, отражающее баланс между precision и recall.

AUC = 0.802: Немного ниже, чем у логистической регрессии, но все равно на высоком уровне.

Вывод: Random Forest показывает результаты, практически идентичные Logistic Regression по общей точности и сбалансированным метрикам, с незначительными преимуществами по precision, но немного уступая по recall и AUC.

XGBoost

Accuracy = 0.747: Наивысшая точность среди рассмотренных моделей — 74,7% объектов классифицируются правильно.

Precision = 0.755: Более высокая точность, чем в предыдущих моделях (75,5% предсказанных положительных являются истинными положительными).

Recall = 0.732: Остается на уровне Logistic Regression (73,2%), что указывает на схожую способность обнаруживать все положительные случаи.

F1 = 0.743: Сбалансированное значение, указывающее на общее улучшение по сравнению с Logistic Regression и Random Forest.

AUC = 0.791: Значение AUC несколько ниже, чем у Logistic Regression и Random Forest, что может говорить о том, что способность модели различать классы при изменении порога оказалась чуть менее выраженной.

Вывод: XGBoost демонстрирует наивысшую точность, precision и F1-score, что указывает на лучшее качество классификации при фиксированном пороге. Однако несколько более низкое AUC говорит о том, что при варьировании порога его способность различать классы немного уступает другим моделям.

LightGBM

Accuracy = 0.711: Общая точность составляет 71,1%, что ниже, чем у остальных моделей.

Precision = 0.716: Значение precision составляет 71,6% — ниже показатели Logistic Regression и Random Forest.

Recall = 0.701: Полнота ниже (70,1%), что указывает на то, что модель пропускает больше реальных положительных случаев.

F1 = 0.708: Итоговый сбалансированный показатель также ниже.

AUC = 0.802: При этом способность различать классы по AUC (0.802) остаётся на уровне Random Forest.

Вывод: Хотя LightGBM показывает конкурентоспособное значение AUC, его значения точности, precision, recall и F1-score оказались ниже, что указывает на более слабую классификационную способность при выбранном пороге.

Общая интерпретация и рекомендации

Сравнение по общей точности и сбалансированным метрикам:

XGBoost лидирует по accuracy (0.747), precision (0.755) и F1 (0.743). Это говорит о том, что при фиксированном пороге XGBoost лучше всего справляется с классификацией объектов.

Logistic Regression и Random Forest показывают очень схожие результаты (accuracy ≈ 0.727), что свидетельствует о стабильности базовой модели и ансамблевого подхода.

LightGBM отстаёт по основным метрикам (accuracy 0.711, F1 0.708), несмотря на хороший AUC, что может означать, что оптимизация порога или гиперпараметров может улучшить его результаты.

Анализ AUC:

Модель Logistic Regression демонстрирует лучший AUC (0.806), что указывает на высокую способность правильно ранжировать объекты по вероятности принадлежности к положительному классу.

Random Forest и LightGBM также показывают высокий AUC (около 0.802), тогда как XGBoost немного отстаёт (0.791). Это может свидетельствовать о том, что, несмотря на высокую классификационную точность XGBoost при выбранном пороге, его способность различать классы при изменении порога несколько ниже.

Вывод по выбору модели:

Если основная цель — максимизировать точность при фиксированном пороге, то XGBoost представляется предпочтительным выбором.

Если важна интерпретация вероятностей и способность различать классы (например, для построения ROC-кривой или дальнейшей калибровки), то Logistic Regression или Random Forest могут быть более выгодными.

LightGBM в данном случае показывает несколько менее удовлетворительные результаты, поэтому для его применения потребуется дополнительная настройка.

Таким образом, все модели демонстрируют свою силу в зависимости от выбранных приоритетов (точность vs. дискриминация по AUC), и окончательный выбор модели может зависеть от специфики задачи, требований к интерпретируемости и дальнейшего использования предсказанных вероятностей.

6.Классификация: CC50 > медианы

Logistic Regression

Accuracy = 0.675: Модель правильно классифицирует 67.5% объектов.

Precision = 0.663: Из всех предсказанных положительных случаев 66.3% действительно истинны, что указывает на умеренное число ложноположительных ошибок.

Recall = 0.711: Модель обнаруживает 71.1% реальных положительных случаев, что демонстрирует её способность выявлять положительные примеры.

F1 = 0.687: Сбалансированный показатель между точностью и полнотой, подтверждающий среднее качество классификации.

AUC = 0.777: Значение свидетельствует о приемлемой способности различать классы, хотя уступает ансамблевым методам.

Вывод: Logistic Regression демонстрирует базовый уровень качества, показывая умеренные значения по всем метрикам, что может служить отправной точкой для сравнения с более сложными алгоритмами.

Random Forest

Accuracy = 0.727: Доля верных предсказаний выросла до 72.7%, что свидетельствует о лучшем общем качестве.

Precision = 0.750: Из предсказанных положительных случаев 75.0% верно определены, что говорит о снижении ложноположительных ошибок.

Recall = 0.680: Полнота несколько снизилась до 68.0% — модель меньше находит реальные положительные, чем Logistic Regression, но это может указывать на более консервативный подход к классификации.

F1 = 0.714: Улучшенное гармоническое среднее между precision и recall подтверждает общий рост качества модели.

AUC = 0.805: Увеличенное значение AUC демонстрирует лучшую способность различать классы.

Вывод: Random Forest показывает улучшение по большинству метрик по сравнению с Logistic Regression, особенно в части точности и способности разделять классы. Немного более низкое значение recall может быть компенсировано высоким precision.

XGBoost

Accuracy = 0.747: Самая высокая общая точность — модель правильно классифицирует 74.7% объектов.

Precision = 0.755: Очень высокий показатель точности (75.5%) свидетельствует о минимуме ложноположительных предсказаний.

Recall = 0.732: Модель обнаруживает 73.2% истинных положительных примеров, что является хорошим балансом с точностью.

F1 = 0.743: Крайне сбалансированное сочетание precision и recall подтверждает высокое качество классификации.

AUC = 0.817: Наивысший показатель AUC говорит о самой сильной способности выделять и различать классы.

Вывод: XGBoost демонстрирует наилучшие результаты по большинству метрик, что делает его оптимальным выбором для задачи классификации. Он комбинирует высокую точность, хорошую полноту и отличную способность разделять классы.

LightGBM

Accuracy = 0.732: Общая точность составляет 73.2% — немного ниже, чем у XGBoost, но всё равно значительно выше, чем у Logistic Regression.

Precision = 0.742: Точность классификации положительного класса на уровне 74.2% говорит о хорошем управлении ложноположительными случаями.

Recall = 0.711: Модель обнаруживает 71.1% реальных позитивов, что является удовлетворительным результатом.

F1 = 0.726: Баланс между precision и recall подтверждается F1-мерой на уровне 72.6%.

AUC = 0.811: Значение AUC указывает на высокую способность разделять классы, хотя немного уступает XGBoost.

Вывод: LightGBM демонстрирует результаты, сопоставимые с Random Forest и XGBoost, с небольшим снижением по некоторым метрикам по сравнению с XGBoost, однако оставаясь на высоком уровне по всем параметрам.

Общие выводы

Сравнение с Logistic Regression:  
Ансамблевые модели (Random Forest, XGBoost и LightGBM) значительно превосходят Logistic Regression по всем ключевым метрикам. Это свидетельствует о том, что в задачах с более сложными зависимостями между признаками ансамблевые методы способны уловить больше информации, чем простая линейная модель.

Лучшая модель:  
XGBoost показывает наилучшие показатели — наивысшую точность, превосходные balance метрики (precision, recall, F1) и максимальное значение AUC. Это делает его наиболее эффективным инструментом для классификации в данном случае.

Выбор модели:  
Хотя XGBoost лидирует, все ансамблевые модели показывают близкие результаты. Выбор между ними может зависеть от специфики задачи, требований к скорости обучения/прогнозирования и интерпретируемости модели. Random Forest и LightGBM могут быть предпочтительны в случаях, когда важна более быстрая настройка или меньшая вычислительная сложность.

Таким образом, применение ансамблевых методов (Random Forest, XGBoost, LightGBM) позволяет существенно улучшить качество классификации по сравнению с Logistic Regression, а особенно заметным результатом обладает XGBoost, демонстрируя наилучший баланс между всеми метриками.

7.Классификация: SI > медианы

Logistic Regression

Accuracy = 0.620: Модель правильно классифицирует около 62% всех наблюдений.

Precision = 0.622: Из предсказанных положительных случаев примерно 62.2% действительно являются положительными.

Recall = 0.610: Модель находит 61.0% всех реальных положительных примеров.

F1 = 0.616: Гармоническое среднее между precision и recall, что отражает сбалансированное качество по этим показателям.

AUC = 0.665: Относительно умеренное значение AUC указывает на то, что модель имеет среднюю способность различать классы при варьировании порога.

Интерпретация: Logistic Regression демонстрирует базовый уровень результатов. При фиксированном пороге точность около 62% и сбалансированные показатели (precision, recall и F1) свидетельствуют об отсутствии явного смещения в сторону либо слишком агрессивного, либо слишком консервативного классифицирования. Однако AUC указывает на то, что различение классов можно улучшить.

Random Forest

Accuracy = 0.630: Немного лучше, чем у логистической регрессии – около 63% верных предсказаний.

Precision = 0.635: Чуть повышенное значение точности говорит о том, что из предсказанных положительных случаев 63.5% действительно положительны.

Recall = 0.610: Полнота остаётся на уровне Logistic Regression (61.0%), что свидетельствует о схожей способности обнаружения положительных примеров.

F1 = 0.622: Объединяет показатели precision и recall, показывая незначительное улучшение по сравнению с логистической регрессией.

AUC = 0.675: Незначительно выше, чем у LR, что указывает на улучшенную (но всё ещё среднюю) способность различать классы.

Интерпретация: Random Forest показывает небольшое улучшение по Accuracy, Precision и AUC по сравнению с LR. Это может свидетельствовать о том, что ансамблевый подход, основанный на объединении множества решающих деревьев, позволяет немного лучше уловить зависимости в данных и уменьшить количество ложноположительных случаев.

XGBoost

Accuracy = 0.610: Общая точность несколько ниже – около 61% правильных предсказаний.

Precision = 0.615: Из предсказанных положительных около 61.5% действительно положительные.

Recall = 0.590: Полнота несколько снизилась – модель обнаруживает 59% положительных примеров, что ниже по сравнению с LR и RF.

F1 = 0.602: Гармоническое среднее также ниже, что отражает некоторое снижение общей сбалансированности.

AUC = 0.649: Значение AUC ниже, чем у остальных рассмотренных моделей, что говорит о меньшей дискриминирующей способности при изменении порога.

Интерпретация: XGBoost в данном случае показывает несколько худшие показатели по всем метрикам. Более низкое значение recall и AUC может указывать на то, что модель либо недостаточно хорошо подобрана с точки зрения гиперпараметров, либо специфика данных не позволяет ей раскрыть свой потенциал. В этом случае модель склонна пропускать больше положительных примеров и имеет ограниченную способность различать классы.

LightGBM

Accuracy = 0.635: Достигается наивысшая точность среди моделей – 63.5%.

Precision = 0.642: Наивысшее значение precision (64.2%) – из положительных предсказаний почти два из трёх окажутся истинными.

Recall = 0.610: Полнота остаётся на уровне 61%, как у LR и RF.

F1 = 0.626: Сбалансированное значение, немного выше, чем у логистической регрессии и XGBoost.

AUC = 0.665: Значение AUC аналогично Logistic Regression, что указывает на умеренную способность различать классы.

Интерпретация: LightGBM показывает лучшие результаты по общим метрикам Accuracy, Precision и F1, что делает её предпочтительным вариантом среди рассмотренных моделей с точки зрения классификации при фиксированном пороге. При этом показатель recall остаётся на среднем уровне, а AUC соответствует LR, то есть способность модели различать классы могла бы быть улучшена.

Общие выводы и сравнение моделей:

LightGBM и Random Forest демонстрируют чуть лучшие показатели по Accuracy и Precision по сравнению с Logistic Regression, что указывает на небольшое преимущество ансамблевых методов в данной задаче.

XGBoost показывает худшие результаты по большинству метрик (Accuracy, Recall, F1 и AUC), что может указывать на неподходящую настройку гиперпараметров или специфику данных, где данный метод менее эффективен.

Метрики всех моделей достаточно близки друг к другу, что может свидетельствовать о том, что предсказательная способность ограничена особенностями данных или выбором признаков.

AUC и дискриминация классов:

Значения AUC варьируются от 0.649 до 0.675, что говорит об умеренной способности моделей правильно ранжировать объекты по вероятности принадлежности к положительному классу.

Улучшить AUC можно, проведя дополнительное исследование предобработки данных, инженерии признаков или оптимизации гиперпараметров.

Классификация: SI > 8

1. Logistic Regression

Accuracy = 0.711:  
Модель правильно классифицирует 71,1% объектов, что является наивысшим показателем среди представленных моделей.

Precision = 0.600:  
Из всех объектов, предсказанных как положительные, 60,0% действительно положительные. Это говорит о некотором количестве ложноположительных случаев.

Recall = 0.565:  
Модель обнаруживает 56,5% всех реальных положительных случаев, что означает неплохую, но не выдающуюся способность захватывать все нужные объекты.

F1 = 0.582:  
Баланс между precision и recall держится на среднем уровне — примерно 58,2%.

AUC = 0.703:  
Это значение свидетельствует об умеренной способности логистической регрессии различать классы при изменении порога.

Интерпретация:  
Логистическая регрессия показывает наивысшую accuracy, что может свидетельствовать о том, что её решения при фиксированном пороге дают достаточно верные результаты. Однако относительная невысокая полнота (Recall) и AUC указывают, что при изменении порога модель не так хорошо различает положительные и отрицательные случаи.

2. Random Forest

Accuracy = 0.691:  
Общая точность чуть ниже — 69,1%, что немного уступает Logistic Regression.

Precision = 0.579:  
Из предсказанных положительных объектов 57,9% являются истинно положительными, что немного хуже логистической регрессии.

Recall = 0.478:  
Модель обнаруживает лишь 47,8% реальных положительных случаев — значение заметно ниже, что указывает на проблематику в находке всех позитивов.

F1 = 0.524:  
Гармоническое среднее между precision и recall также ниже, чем у Logistic Regression, отражая общую потерю в балансе между точностью и полнотой.

AUC = 0.731:  
Несмотря на несколько худшие показатели при фиксированном пороге, AUC выше (0.731). Это говорит о лучшей способности модели ранжировать объекты, даже если выбранный порог не оптимален.

Интерпретация:  
Random Forest работает несколько осторожнее, что выражается в меньшей полноте (то есть модель с большей вероятностью пропускает реальные положительные примеры), но при этом способность различать классы (AUC) улучшена по сравнению с Logistic Regression. Это может указывать на потенциал модели при дополнительной настройке порогов или весов.

3. XGBoost

Accuracy = 0.696:  
Общая точность составляет 69,6%, что немного ниже показателя Logistic Regression, но сопоставимо с Random Forest.

Precision = 0.566:  
Из предсказанных положительных только 56,6% являются истинными положительными, то есть модель склонна выдавать больше ложноположительных результатов.

Recall = 0.623:  
При этом полнота равна 62,3% — самое высокое значение среди моделей, что свидетельствует о лучшем обнаружении всех позитивных случаев.

F1 = 0.593:  
Баланс между precision и recall чуть улучшен (59,3%), что говорит о том, что модель работает с небольшим перекосом в сторону повышения обнаружения позитивов за счёт точности.

AUC = 0.735:  
Это один из лучших показателей AUC, указывающий на высокую способность модели различать классы по вероятностной шкале.

Интерпретация:  
XGBoost демонстрирует лучший recall, то есть успешно находит больше истинных положительных примеров, что может быть особенно актуально в задачах, где критично не пропустить позитивы. Несмотря на несколько более низкую precision, AUC остаётся высоким, что говорит о хорошем ранжировании объектов по принадлежности к классам. Таким образом, модель эффективна для задач, где важна полнота, даже если это идет в ущерб точности предсказаний.

4. LightGBM

Accuracy = 0.675:  
Общая точность равна 67,5%, что является самым низким значением среди рассмотренных моделей.

Precision = 0.548:  
Из предсказанных положительных 54,8% действительно положительные – значение precision ниже, чем у остальных моделей.

Recall = 0.493:  
Полнота составляет 49,3%, что сигнализирует о том, что модель пропускает более половины реальных положительных примеров.

F1 = 0.519:  
Сбалансированный показатель между precision и recall также самый низкий — всего 51,9%.

AUC = 0.736:  
Интересно, что AUC LightGBM оказывается самым высоким (или сопоставимым с лучшими) — 0.736, что говорит о том, что при варьировании порога модель способна хорошо разделять классы. Однако при фиксированном пороге ее предсказания оказываются менее точными.

Интерпретация:  
Несмотря на лучший показатель AUC, оценка по фиксированному порогу (accuracy, precision, recall, F1) для LightGBM оказалась самой низкой. Это может означать, что оптимальный порог для принятия решения не был подобран, либо модель требует дополнительной калибровки. Высокий AUC указывает, что LightGBM имеет потенциал для различения классов, если правильно настроить пороговое значение.

Итоговое сравнение и рекомендации

Фиксированный порог vs. способность ранжирования:

Logistic Regression показывает наивысшую accuracy (0.711) и хорошие сбалансированные показатели (precision = 0.600, recall = 0.565, F1 = 0.582), но имеет самый низкий AUC (0.703).

Random Forest уступает Logistic Regression по accuracy, precision и recall, но выигрывает по AUC (0.731), что свидетельствует о потенциале модели при оптимальном выборе порога.

XGBoost демонстрирует самый высокий recall (0.623) и один из лучших показателей AUC (0.735). Это может быть полезно, если задача требует обнаружения всех позитивных случаев, даже если точность предсказаний немного ниже.

LightGBM показывает худшие значения по пороговым метрикам (accuracy, precision, recall, F1), но имеет самый высокий AUC (0.736). Это указывает на то, что LightGBM хорошо различает классы в ранжировании, однако оптимизация порога или калибровка вероятностей может повысить её практическую эффективность.

Выбор модели зависит от задачи:  
Если требуется максимальная точность фиксированного порога, то Logistic Regression оказывается наиболее стабильной.  
Если важнее не пропустить позитивы, XGBoost с его высоким recall является предпочтительным выбором, несмотря на снижение precision.  
Если планируется дальнейшая настройка порога для повышения разделительной способности, модели с высоким AUC (Random Forest и LightGBM) могут дать преимущества после дополнительной оптимизации.

Таким образом, каждая модель имеет свои сильные и слабые стороны. Итоговый выбор модели следует делать исходя из специфики задачи и требований: например, снижать вероятность пропуска важных случаев (Recall) или минимизировать ложноположительные предсказания (Precision), а также учитывать возможность дальнейшей настройки порогов, опираясь на значение AUC.