Национальный исследовательский ядерный университет «МИФИ»
Дисциплина «Классическое машинное обучение»
Пояснительная записка к курсовой работе

Работу выполнил:
Лудков Андрей Сергеевич
Группа М24-525

Оглавление

Введение

Процесс разработки новейших лекарственных препаратов всегда сопряжен с внушительными материальными затратами как на закупку необходимых компонентов, так и на проведение соответствующих исследований. Одним из подходов по оптимизации указанного процесса могут являться удачно подобранные алгоритмы машинного обучения, способные без использования реальных реактивов спрогнозировать эффективность тех или иных химических соединений. Для данного эксперимента достаточно обладать исчерпывающими данными о потенциальных компонентах.

Несомненно, качественная и экспертная оценки фармакологических характеристик соединений невозможна без использования в процессе исследования методов химоинформатики.

В настоящей работе уже представлены конфиденциальные данные о 1000 химических соединениях с указанием их эффективности против вируса гриппа. Данные разделены на структурные, электронные и топологические дескрипторы. Эффективность возможных соединений характеризуется следующими параметрами:

- IC50 – активность;
- CC50 – токсичность;
- SI – селективность.

В рамках курсовой работы необходимо выполнить:

1. Исследовательский анализ данных (EDA);
2. Создание регрессионных моделей:
   - регрессия для IC50;
   - регрессия для CC50;
   - регрессия для SI.
3. Создание моделей для классификации по следующим признакам:
   - превышает ли значение IC50 медианное значение выборки;
   - превышает ли значение CC50 медианное значение выборки;
   - превышает ли значение SI медианное значение выборки;
   - превышает ли значение SI значение 8.
4. Сравнить полученные модели в рамках отдельных подзадач, осуществить оптимизацию гиперпараметров, оценить предсказательную способность и составить рекомендации по дальнейшему улучшению.

1 Разведывательный анализ данных (EDA)

В рамках разработки моделей машинного обучения всегда необходимо провести тщательный разведывательный анализ данных с целью повышения предсказательных способностей будущих моделей.

В настоящей работе оценены распределения целевых показателей (IC50, CC50, SI), произведен анализ пропусков и выбросов, осуществлен корреляционный анализ данных и выявлены наиболее значимые признаки.

1.1 Описание данных

Таблица 1 – Описание признаков

Признак	Описание
Общие молекулярные дескрипторы	
MolWt	Молекулярная масса
HeavyAtomCount	Количество тяжелых атомов (без H)
NumValenceElectrons	Валентные электроны
NumRadicalElectrons	Неспаренные электроны
FractionCSP3	Доля sp3-гибридизованных атомов C
TPSA	Топологическая полярная поверхность (проницаемость через мембраны)
LabuteASA	Доступная поверхность по Labute (взаимодействие с растворителем)
QED	Оценка «лекарственности» (комплексный показатель)
MolLogP	Гидрофобность (logP)
MolMR	Молекулярная рефрактивность (поляризуемость)
Электронные дескрипторы	
Max/MinPartialCharge	Экстремальные значения частичных зарядов
PEOE_VSA	Распределение зарядов (метод PEOE)
Estate_VSA	Зарядовое состояние и топология
Max/MinEStateIndex	Индексы электротопологического состояния
Топологические дескрипторы	
Chi0-Chi4v	Индексы связности (топология, разветвление)
Kappa1-Kappa3	Индексы формы и компактности
HallKierAlpha	Стерическая насыщенность
BalabanJ	Связность и цикличность (разветвлённость)
Ipc, AvgIpc, BertzCT	Информационные индексы сложности структуры
BCUT-дескрипторы	
BCUT2D_MW	Молекулярная масса (высокая/низкая)
BCUT2D_CHG	Заряд (высокий/низкий)
BCUT2D_LOGP	Гидрофобность (высокая/низкая)
BCUT2D_MR	Рефрактивность (высокая/низкая)
VSA-дескрипторы	
SMR_VSA1-10	Молекулярная рефрактивность по диапазонам
SlogP_VSA1-12	Гидрофобность по участкам
Estate_VSA1-10	Электротопология по поверхности
PEOE_VSA1-14	Частичные заряды по диапазонам
Отпечатки (Morgan fingerprints)	
FPDensityMorgan1-3	Плотность структурных фрагментов (радиусы 1, 2, 3)
Фрагментные дескрипторы	
fr_phenol, fr_Ar_OH	Фенолы
fr_NH2, fr_aniline	Амины
fr_azide, fr_azo	Азосоединения
fr_halogen, fr_alkyl_halide	Галогены

Признак	Описание
fr_barbitur	Барбитураты
fr_nitro, fr_nitro_atom	Нитро-соединения
fr_benzene, fr_pyridine, fr_furan	Кольца
Структурные количественные дескрипторы	
NumHAcceptors/ NumHDonors	Акцепторы/доноры водородных связей
NumRotatableBonds	Вращающиеся связи
NumAromatic/ Aliphatic/SaturatedRings	Типы колец
NumHeteroatoms	Количество гетероатомов
RingCount	Общее число колец

1.2 Первоначальный анализ данных

В рамках базового анализа была исследована структура данных, наличие пропусков и дубликатов.

Были выявлены пропуски в следующих столбцах (были удалены из датасета):

```
MaxPartialCharge       3
MinPartialCharge       3
MaxAbsPartialCharge    3
MinAbsPartialCharge    3
BCUT2D_MWHI            3
BCUT2D_MWLOW           3
BCUT2D_CHGHI           3
BCUT2D_CHGLO           3
BCUT2D_LOGPHI          3
BCUT2D_LOGPLOW         3
BCUT2D_MRHI            3
BCUT2D_MRLOW           3
dtype: int64
```

Рисунок 1.1 – Признаки с пропусками

После выявления и удаления дубликатов общее число строк сократилось с 1001 до 966.

1.3 Распределение целевых показателей

Графики распределения целевых переменных приведены на рисунке 2.

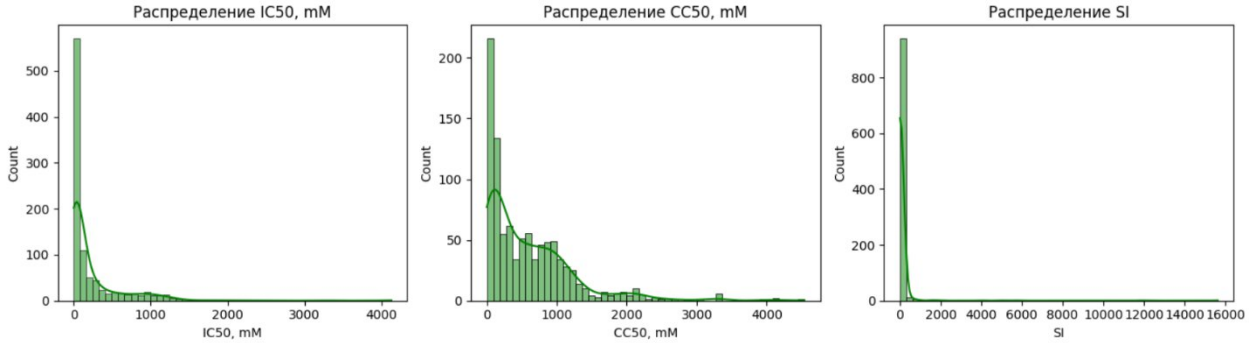

Рисунок 1.2 – Распределение целевых переменных

Полученные графики свидетельствуют об явно выраженном левоассиметричном распределении, что характеризует высокое число выбросов в целевых переменных.

Для улучшения визуализации используем логарифмическое преобразование данных (по основанию 10). Результаты представлены на рисунке 1.3.

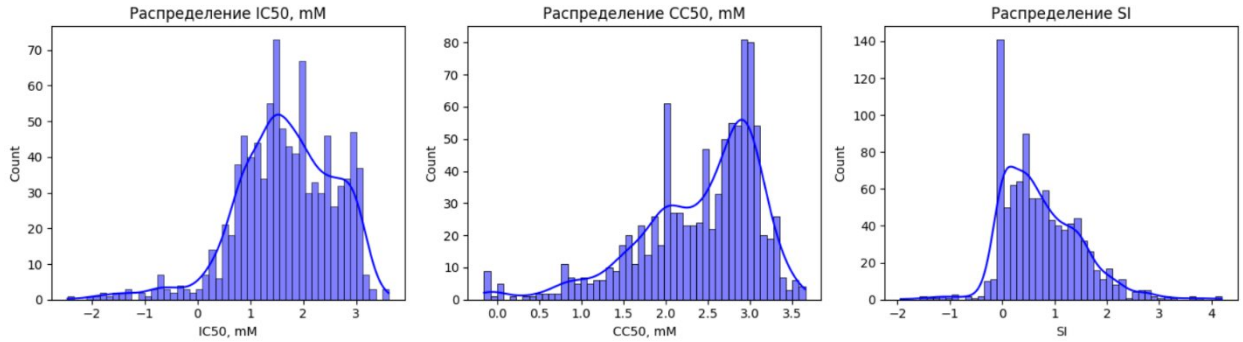

Рисунок 1.3 – Распределение логарифмированных целевых переменных

Стоит обратить внимание на существенное количество значений около нуля в распределении SI. В дальнейшем будет применено отсеивание признаков по правилу трех сигм, что позволит увеличить будущие метрики моделей.

Для визуализации выбросов были построены графики, представленные на рисунке 1.4.

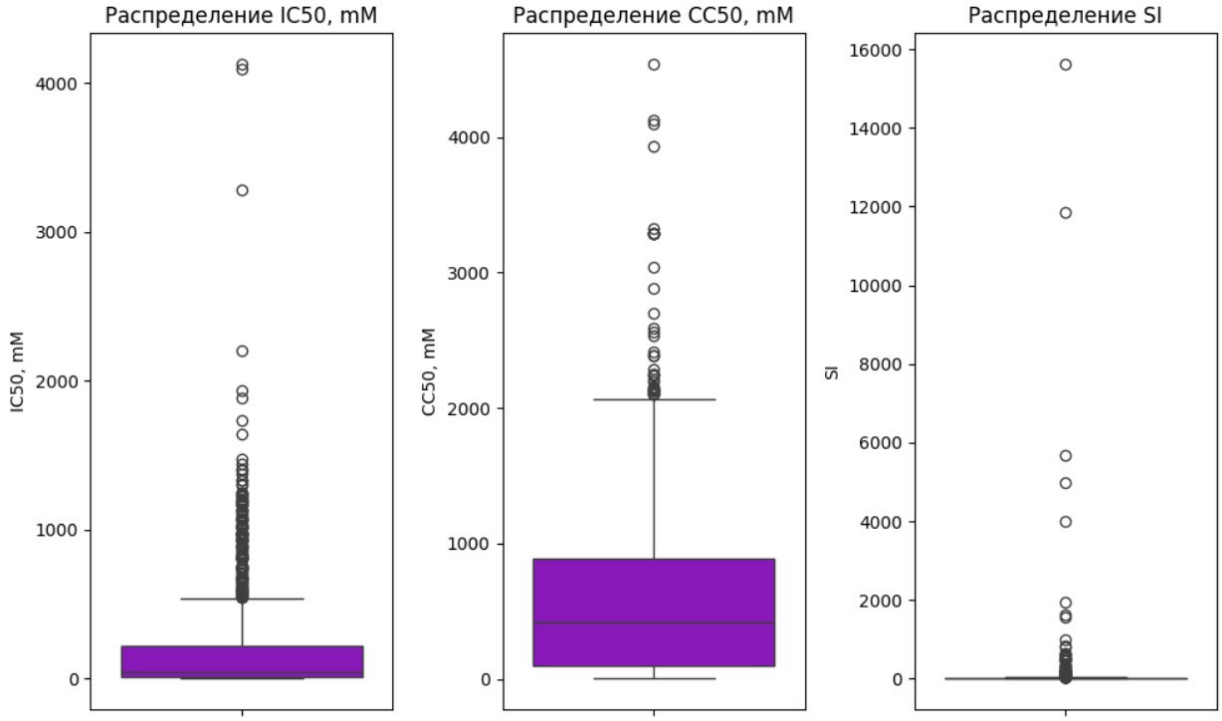

Рисунок 1.4 – Визуализация выбросов целевых переменных

1.4 Корреляционный анализ данных

В рамках настоящего EDA был проведен корреляционный анализ данных для установления линейных зависимостей как между целевыми переменными и парами признаков, так и между признаками с целевыми переменными.

Корреляция признаков и целевых переменных необходима для определения значимости тех или иных признаков и, как следствие, разумному уменьшению объема данных.

Высокая линейная зависимость (более 0,9) между признаками и целевыми переменными не выявлена.

```
Признак: IC50, mM:
Корреляция > 0.9 с IC50, mM: []

Признак: CC50, mM:
Корреляция > 0.9 с CC50, mM: []

Признак: SI:
Корреляция > 0.9 с SI: []
```

Рисунок 1.5 – Результаты корреляционного анализа

Было найдено существенное количество признаков с минимальной линейной зависимостью с целевыми переменными (рисунок 1.6).

```
Признак: IC50, mM:
Корреляция < 0.01 с IC50, mM: Index(['fr_thiophene', 'fr_oxime', 'fr_oxazole', 'fr_nitro_arom_nonortho',
       'fr_nitro_arom', 'fr_morpholine', 'fr_methoxy', 'fr_aryl_methyl',
       'fr_amide', 'fr_aldehyde', 'fr_NH1', 'fr_Al_OH_noTert', 'TPSA',
       'SlogP_VSA2', 'PEOE_VSA9', 'NumHAcceptors'],
      dtype='object')

Признак: CC50, mM:
Корреляция < 0.01 с CC50, mM: Index(['fr_sulfone', 'fr_piperdine', 'fr_nitro_arom_nonortho', 'fr_nitro_arom',
       'fr_imidazole', 'fr_hdrzine', 'fr_azo', 'fr_aldehyde', 'SMR_VSA3', 'SI',
       'PEOE_VSA4', 'PEOE_VSA13', 'NumSaturatedCarbocycles',
       'NumAromaticHeterocycles', 'EState_VSA11'],
      dtype='object')

Признак: SI:
Корреляция < 0.01 с SI: Index(['fr_urea', 'fr_unbrch_alkane', 'fr_tetrazole', 'fr_term_acetylene',
       'fr_sulfone', 'fr_priamide', 'fr_piperzine', 'fr_oxime', 'fr_oxazole',
       'fr_nitro_arom_nonortho', 'fr_nitrile', 'fr_hdrzine', 'fr_epoxide',
       'fr_azo', 'fr_amidine', 'fr_aldehyde', 'fr_HOCCN', 'fr_Ar_COO',
       'VSA_EState9', 'VSA_EState7', 'SPS', 'SMR_VSA2', 'SMR_VSA1',
       'PEOE_VSA8', 'NumAliphaticCarbocycles', 'MaxEStateIndex',
       'MaxAbsEStateIndex', 'Kappa1', 'Ipc', 'EState_VSA1', 'CC50, mM'],
      dtype='object')
```

Рисунок 1.6 – Результаты корреляционного анализа

Далее была оценена линейная зависимость непосредственно между целевыми переменными (рисунок 1.7).

Рисунок 1.7 – Корреляция между целевыми переменными

Полученная матрица свидетельствует о наличии приемлемой линейной зависимости между IC50 и CC50, что указывает на связь между токсичностью и активностью химических соединений.

Непосредственно ключевая корреляционная матрица, охватывающая нецелевые признаки, представлена на рисунке 1.8.

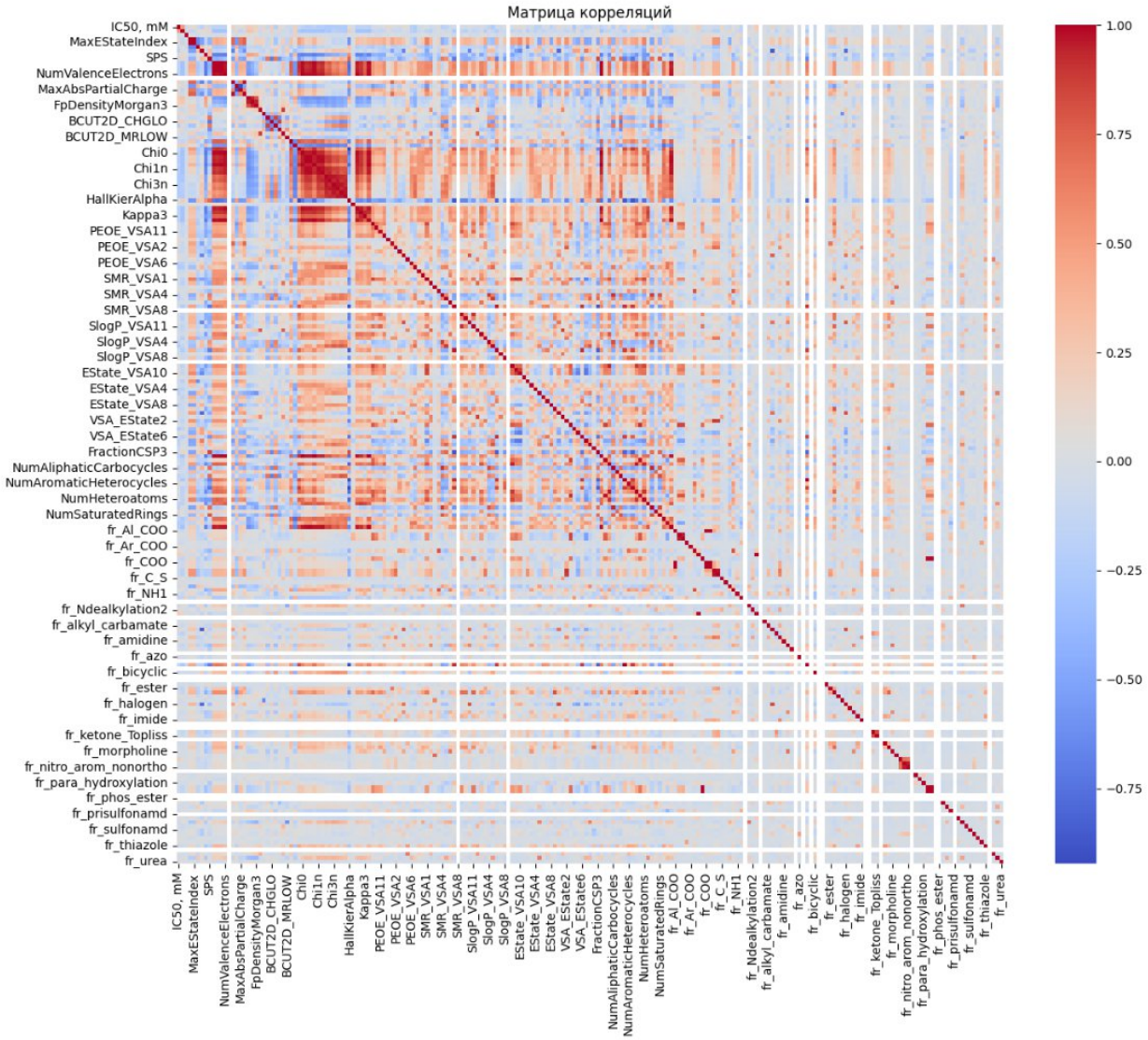

Рисунок 1.8 – Корреляционная матрица для нецелевых признаков

Ввиду большого количества признаков достаточно сложно осуществить подробный анализ полученной матрицы. Однако и в таком виде она явно указывает на наличие существенных линейных зависимостей между нецелевыми показателями.

Был проведен дополнительный попарный анализ признаков. В результате были получены 154 пары столбцов, коэффициенты линейной зависимости которых превышали 0,9:

1. MaxAbsEStateIndex и MaxEStateIndex: 1.000
2. MolWt и HeavyAtomMolWt: 0.997
3. MolWt и ExactMolWt: 1.000
4. MolWt и NumValenceElectrons: 0.981
5. MolWt и BertzCT: 0.902
6. MolWt и Chi0: 0.987
7. MolWt и Chi0n: 0.935
8. MolWt и Chi0v: 0.954

9. MolWt и Chi1: 0.987
10. MolWt и Chi1n: 0.905
11. MolWt и Chi1v: 0.927
12. MolWt и Kappa1: 0.960
13. MolWt и Kappa2: 0.907
14. MolWt и LabuteASA: 0.988
15. MolWt и HeavyAtomCount: 0.989
16. MolWt и MolMR: 0.957
17. HeavyAtomMolWt и ExactMolWt: 0.997
18. HeavyAtomMolWt и NumValenceElectrons: 0.966
19. HeavyAtomMolWt и BertzCT: 0.924
20. HeavyAtomMolWt и Chi0: 0.978
21. HeavyAtomMolWt и Chi0n: 0.906
22. HeavyAtomMolWt и Chi0v: 0.929
23. HeavyAtomMolWt и Chi1: 0.982
24. HeavyAtomMolWt и Kappa1: 0.940
25. HeavyAtomMolWt и LabuteASA: 0.977
26. HeavyAtomMolWt и HeavyAtomCount: 0.983
27. HeavyAtomMolWt и MolMR: 0.937
28. ExactMolWt и NumValenceElectrons: 0.981
29. ExactMolWt и BertzCT: 0.902
30. ExactMolWt и Chi0: 0.987
31. ExactMolWt и Chi0n: 0.935
32. ExactMolWt и Chi0v: 0.954
33. ExactMolWt и Chi1: 0.987
34. ExactMolWt и Chi1n: 0.905
35. ExactMolWt и Chi1v: 0.927
36. ExactMolWt и Kappa1: 0.960
37. ExactMolWt и Kappa2: 0.907
38. ExactMolWt и LabuteASA: 0.989
39. ExactMolWt и HeavyAtomCount: 0.989
40. ExactMolWt и MolMR: 0.957
41. NumValenceElectrons и Chi0: 0.995
42. NumValenceElectrons и Chi0n: 0.976
43. NumValenceElectrons и Chi0v: 0.975
44. NumValenceElectrons и Chi1: 0.985
45. NumValenceElectrons и Chi1n: 0.953
46. NumValenceElectrons и Chi1v: 0.950
47. NumValenceElectrons и Kappa1: 0.987
48. NumValenceElectrons и Kappa2: 0.928
49. NumValenceElectrons и LabuteASA: 0.991
50. NumValenceElectrons и HeavyAtomCount: 0.991
51. NumValenceElectrons и MolMR: 0.968
52. MaxPartialCharge и MinAbsPartialCharge: 0.974
53. MinPartialCharge и MaxAbsPartialCharge -0.922
54. FpDensityMorgan1 и FpDensityMorgan2: 0.948
55. FpDensityMorgan2 и FpDensityMorgan3: 0.940
56. BertzCT и Chi1: 0.916

57. BertzCT и HallKierAlpha -0.904
58. BertzCT и HeavyAtomCount: 0.907
59. Chi0 и Chi0n: 0.960
60. Chi0 и Chi0v: 0.960
61. Chi0 и Chi1: 0.991
62. Chi0 и Chi1n: 0.927
63. Chi0 и Chi1v: 0.928
64. Chi0 и Kappa1: 0.980
65. Chi0 и Kappa2: 0.923
66. Chi0 и LabuteASA: 0.990
67. Chi0 и HeavyAtomCount: 0.996
68. Chi0 и MolMR: 0.961
69. Chi0n и Chi0v: 0.993
70. Chi0n и Chi1: 0.948
71. Chi0n и Chi1n: 0.990
72. Chi0n и Chi1v: 0.979
73. Chi0n и Chi2n: 0.903
74. Chi0n и Kappa1: 0.967
75. Chi0n и LabuteASA: 0.972
76. Chi0n и HeavyAtomCount: 0.955
77. Chi0n и MolMR: 0.980
78. Chi0v и Chi1: 0.954
79. Chi0v и Chi1n: 0.982
80. Chi0v и Chi1v: 0.989
81. Chi0v и Chi2v: 0.907
82. Chi0v и Kappa1: 0.964
83. Chi0v и LabuteASA: 0.981
84. Chi0v и HeavyAtomCount: 0.959
85. Chi0v и MolMR: 0.989
86. Chi1 и Chi1n: 0.922
87. Chi1 и Chi1v: 0.927
88. Chi1 и Kappa1: 0.955
89. Chi1 и Kappa2: 0.913
90. Chi1 и LabuteASA: 0.993
91. Chi1 и HeavyAtomCount: 0.999
92. Chi1 и MolMR: 0.968
93. Chi1n и Chi1v: 0.984
94. Chi1n и Chi2n: 0.935
95. Chi1n и Chi2v: 0.924
96. Chi1n и Kappa1: 0.937
97. Chi1n и LabuteASA: 0.951
98. Chi1n и HeavyAtomCount: 0.928
99. Chi1n и MolMR: 0.969
100. Chi1v и Chi2n: 0.910
101. Chi1v и Chi2v: 0.940
102. Chi1v и Kappa1: 0.934
103. Chi1v и LabuteASA: 0.960
104. Chi1v и HeavyAtomCount: 0.932

105. Chi1v и MolMR: 0.977
106. Chi2n и Chi2v: 0.971
107. Chi2n и Chi3n: 0.967
108. Chi2n и Chi3v: 0.948
109. Chi2n и Chi4n: 0.932
110. Chi2n и Chi4v: 0.906
111. Chi2v и Chi3n: 0.930
112. Chi2v и Chi3v: 0.965
113. Chi2v и Chi4v: 0.930
114. Chi3n и Chi3v: 0.972
115. Chi3n и Chi4n: 0.965
116. Chi3n и Chi4v: 0.930
117. Chi3v и Chi4n: 0.938
118. Chi3v и Chi4v: 0.965
119. Chi4n и Chi4v: 0.966
120. Kappa1 и Kappa2: 0.956
121. Kappa1 и LabuteASA: 0.968
122. Kappa1 и HeavyAtomCount: 0.964
123. Kappa1 и MolMR: 0.950
124. Kappa2 и Kappa3: 0.936
125. Kappa2 и LabuteASA: 0.915
126. Kappa2 и HeavyAtomCount: 0.913
127. Kappa2 и MolMR: 0.905
128. LabuteASA и HeavyAtomCount: 0.994
129. LabuteASA и MolMR: 0.986
130. SMR_VSA7 и SlogP_VSA6: 0.959
131. SMR_VSA7 и VSA_EState6: 0.902
132. SMR_VSA7 и NumAromaticCarbocycles: 0.909
133. SMR_VSA7 и fr_benzene: 0.909
134. SMR_VSA9 и SlogP_VSA11: 0.913
135. SlogP_VSA6 и VSA_EState6: 0.923
136. TPSA и NOCount: 0.936
137. VSA_EState2 и fr_C_O: 0.905
138. VSA_EState3 и NumHDonors: 0.919
139. HeavyAtomCount и MolMR: 0.968
140. NHOHCount и NumHDonors: 0.981
141. NOCount и NumHAcceptors: 0.956
142. NOCount и NumHeteroatoms: 0.923
143. NumAliphaticCarbocycles и NumSaturatedCarbocycles: 0.925
144. NumAromaticCarbocycles и fr_benzene: 1.000
145. fr_Al_COO и fr_COO: 0.990
146. fr_Al_COO и fr_COO2: 0.990
147. fr_Al_OH и fr_Al_OH_noTert: 0.956
148. fr_Ar_NH и fr_Nhpyrrole: 1.000
149. fr_Ar_OH и fr_phenol: 0.991
150. fr_Ar_OH и fr_phenol_noOrthoHbond: 0.991
151. fr_COO и fr_COO2: 1.000
152. fr_C_O и fr_C_O_noCOO: 0.976

153. fr_nitro_arom и fr_nitro_arom_nonortho: 0.957
154. fr_phenol и fr_phenol_noOrthoHbond: 1.000

1.5 Поиск выбросов в данных

Наличие выбросов негативно сказывается на предсказательной способности моделей. Визуализация выбросов в целевых показателях была представлена на рисунке 1.4.

В ходе разведывательного анализа были выявлены выбросы по правилу трех сигм (предельными значениями были взяты 1-й и 3-й квартили).

Таким образом, были получены суммарные значения выбросов для всех столбцов датасета. Количество выбросов в целевых показателях приведено на рисунке 1.9.

```
Outliers (IC50): 140/966
Outliers (CC50): 35/966
Outliers (SI): 119/966
```

Рисунок 1.9 – Количество выбросов в целевых переменных

Дополнительно был проведен аналогичный анализ и для нецелевых признаков:

1. MaxAbsEStateIndex: 60/966
2. MaxEStateIndex: 60/966
3. MinAbsEStateIndex: 21/966
4. MinEStateIndex: 121/966
5. MolWt: 38/966
6. HeavyAtomMolWt: 37/966
7. ExactMolWt: 38/966
8. NumValenceElectrons: 45/966
9. MinPartialCharge: 5/966
10. MaxAbsPartialCharge: 4/966
11. FpDensityMorgan1: 3/966
12. FpDensityMorgan2: 31/966
13. FpDensityMorgan3: 50/966
14. BCUT2D_MWHI: 39/966
15. BCUT2D_MWLOW: 11/966
16. BCUT2D_CHGHI: 28/966
17. BCUT2D_CHGLO: 3/966
18. BCUT2D_LOGPHI: 4/966
19. BCUT2D_LOGPLOW: 37/966
20. BCUT2D_MRHI: 55/966
21. BCUT2D_MRLOW: 69/966
22. BalabanJ: 36/966
23. BertzCT: 34/966
24. Chi0: 34/966
25. Chi0n: 53/966
26. Chi0v: 62/966
27. Chi1: 38/966

28. Chi1n: 46/966
29. Chi1v: 51/966
30. Chi2n: 31/966
31. Chi2v: 24/966
32. Chi3n: 51/966
33. Chi3v: 40/966
34. Chi4n: 50/966
35. Chi4v: 55/966
36. HallKierAlpha: 2/966
37. Ipc: 200/966
38. Kappa1: 42/966
39. Kappa2: 59/966
40. Kappa3: 62/966
41. LabuteASA: 47/966
42. PEOE_VSA1: 40/966
43. PEOE_VSA10: 52/966
44. PEOE_VSA11: 120/966
45. PEOE_VSA12: 69/966
46. PEOE_VSA13: 183/966
47. PEOE_VSA14: 10/966
48. PEOE_VSA2: 24/966
49. PEOE_VSA3: 48/966
50. PEOE_VSA4: 197/966
51. PEOE_VSA5: 163/966
52. PEOE_VSA6: 19/966
53. PEOE_VSA7: 42/966
54. PEOE_VSA8: 25/966
55. PEOE_VSA9: 29/966
56. SMR_VSA1: 20/966
57. SMR_VSA10: 23/966
58. SMR_VSA2: 11/966
59. SMR_VSA3: 11/966
60. SMR_VSA4: 22/966
61. SMR_VSA5: 27/966
62. SMR_VSA6: 48/966
63. SMR_VSA9: 141/966
64. SlogP_VSA1: 8/966
65. SlogP_VSA10: 104/966
66. SlogP_VSA11: 109/966
67. SlogP_VSA12: 12/966
68. SlogP_VSA2: 15/966
69. SlogP_VSA3: 33/966
70. SlogP_VSA4: 22/966
71. SlogP_VSA5: 36/966
72. SlogP_VSA6: 3/966
73. SlogP_VSA7: 55/966
74. SlogP_VSA8: 103/966
75. TPSA: 22/966

76. EState_VSA1: 48/966
77. EState_VSA10: 39/966
78. EState_VSA11: 26/966
79. EState_VSA2: 57/966
80. EState_VSA3: 52/966
81. EState_VSA4: 28/966
82. EState_VSA5: 45/966
83. EState_VSA6: 20/966
84. EState_VSA7: 42/966
85. EState_VSA8: 18/966
86. EState_VSA9: 41/966
87. VSA_EState1: 44/966
88. VSA_EState10: 144/966
89. VSA_EState2: 30/966
90. VSA_EState3: 62/966
91. VSA_EState4: 11/966
92. VSA_EState5: 54/966
93. VSA_EState6: 17/966
94. VSA_EState7: 30/966
95. VSA_EState8: 35/966
96. VSA_EState9: 237/966
97. HeavyAtomCount: 41/966
98. NHOHCount: 23/966
99. NOCount: 25/966
100. NumAliphaticCarbocycles: 15/966
101. NumAliphaticHeterocycles: 1/966
102. NumAliphaticRings: 8/966
103. NumAromaticHeterocycles: 15/966
104. NumAromaticRings: 7/966
105. NumHAcceptors: 96/966
106. NumHDonors: 23/966
107. NumHeteroatoms: 13/966
108. NumRotatableBonds: 44/966
109. NumSaturatedCarbocycles: 2/966
110. NumSaturatedHeterocycles: 11/966
111. NumSaturatedRings: 3/966
112. RingCount: 28/966
113. MolLogP: 61/966
114. MolMR: 38/966
115. fr_Al_COO: 53/966
116. fr_Al_OH: 238/966
117. fr_Al_OH_noTert: 177/966
118. fr_ArN: 14/966
119. fr_Ar_COO: 1/966
120. fr_Ar_N: 87/966
121. fr_Ar_NH: 31/966
122. fr_Ar_OH: 81/966
123. fr_COO: 54/966

124. fr_COO2: 54/966
125. fr_C_O: 73/966
126. fr_C_O_noCOO: 59/966
127. fr_C_S: 42/966
128. fr_HOCCN: 1/966
129. fr_Imine: 133/966
130. fr_NH0: 38/966
131. fr_NH1: 10/966
132. fr_NH2: 119/966
133. fr_Ndealkylation1: 61/966
134. fr_Ndealkylation2: 77/966
135. fr_Nhpyrrole: 31/966
136. fr_aldehyde: 3/966
137. fr_alkyl_carbamate: 12/966
138. fr_alkyl_halide: 129/966
139. fr_allylic_oxid: 206/966
140. fr_amide: 239/966
141. fr_amidine: 8/966
142. fr_aniline: 221/966
143. fr_aryl_methyl: 143/966
144. fr_azo: 7/966
145. fr_bicyclic: 20/966
146. fr_epoxide: 4/966
147. fr_ester: 199/966
148. fr_ether: 42/966
149. fr_furan: 44/966
150. fr_guanido: 4/966
151. fr_halogen: 131/966
152. fr_hdrzine: 3/966
153. fr_hdrzone: 64/966
154. fr_imidazole: 52/966
155. fr_imide: 27/966
156. fr_ketone: 150/966
157. fr_ketone_Topliss: 83/966
158. fr_lactone: 40/966
159. fr_methoxy: 152/966
160. fr_morpholine: 52/966
161. fr_nitrile: 6/966
162. fr_nitro: 24/966
163. fr_nitro_arom: 12/966
164. fr_nitro_arom_nonortho: 11/966
165. fr_oxazole: 4/966
166. fr_oxime: 8/966
167. fr_para_hydroxylation: 143/966
168. fr_phenol: 77/966
169. fr_phenol_noOrthoHbond: 77/966
170. fr_piperdine: 58/966
171. fr_piperzine: 13/966

172. fr_priamide: 23/966
173. fr_pyridine: 25/966
174. fr_quatN: 36/966
175. fr_sulfide: 45/966
176. fr_sulfonamd: 12/966
177. fr_sulfone: 9/966
178. fr_term_acetylene: 1/966
179. fr_tetrazole: 1/966
180. fr_thiazole: 52/966
181. fr_thiophene: 68/966
182. fr_unbrch_alkane: 49/966
183. fr_urea: 7/966

Дополнительно был построен график первых сорока признаков по общему количеству выбросов, представленный на рисунке 1.10.

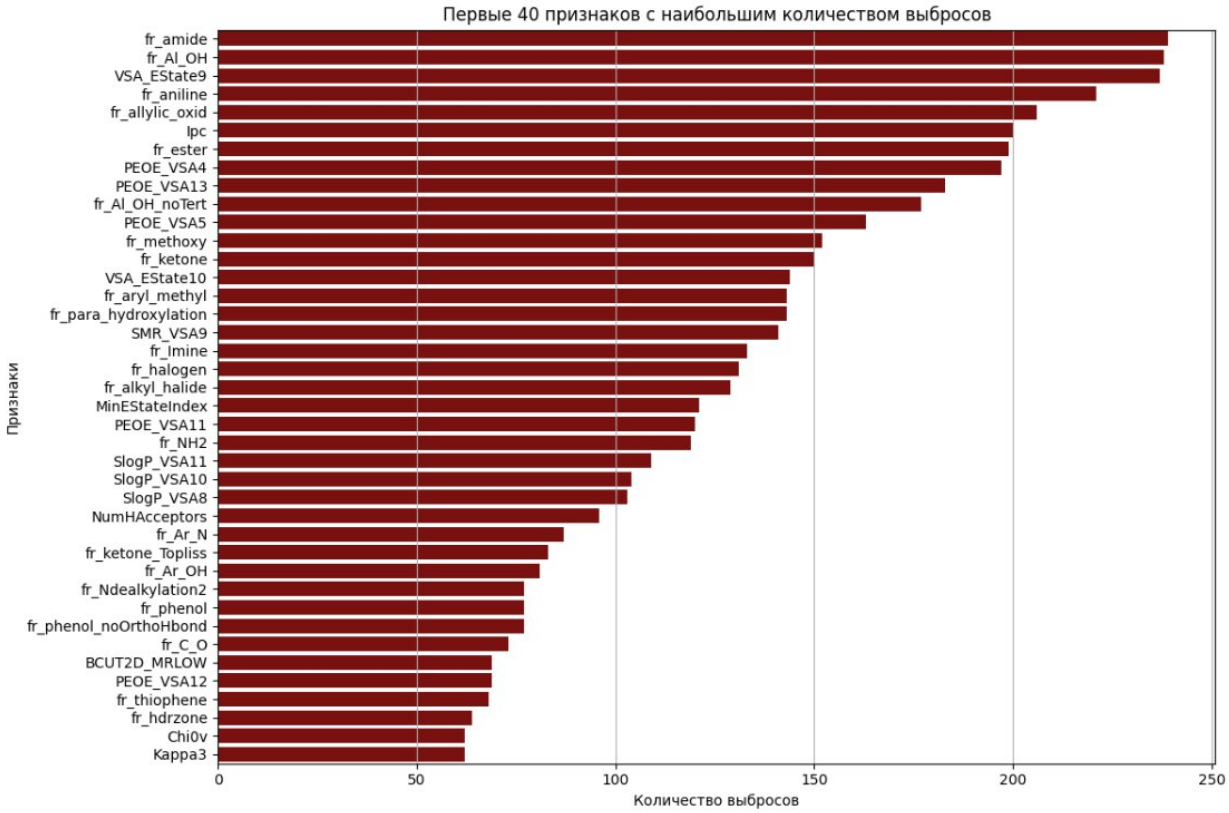

Рисунок 1.10 – Топ-40 признаков с наибольшим числом выбросов

Полученные результаты анализа свидетельствуют о существенном количестве выбросов как в целевых, так и нецелевых признаках.

1.6 Отбор наиболее значимых признаков

В рамках разведывательного анализа данных было произведено сокращение числа признаков датасета посредством следующих показателей:

– нулевая дисперсия;
– корреляционный анализ.

Первоначальное число признаков – 213 (в самом начале был удален столбец Unnamed: 0 вследствие отсутствия значимости).

Сперва были удалены признаки с нулевой дисперсией (приведены на рисунке 1.11).

```
Index(['NumRadicalElectrons', 'SMR_VSA8', 'SlogP_VSA9', 'fr_N_O', 'fr_SH',
       'fr_azide', 'fr_barbitur', 'fr_benzodiazepine', 'fr_diazo',
       'fr_dihydropyridine', 'fr_isocyan', 'fr_isothiocyan', 'fr_lactam',
       'fr_nitroso', 'fr_phos_acid', 'fr_phos_ester', 'fr_prisulfonamd',
       'fr_thiocyan'],
      dtype='object')
(966, 195)
```

Рисунок 1.11 – Признаки с нулевой дисперсией

Общее число столбцов сократилось с 213 до 195.

Следующим шагом были удалены столбцы, имеющие наибольшую линейную зависимость, установленную в результате корреляционного анализа. Из каждой полученной пары таких признаков удалялся правый столбец, что позволило в конечном итоге сократить число столбцов до 148.

1.7 Удаление выбросов и нормализация данных

В рамках настоящего задания были удалены выбросы только в целевых переменных. Опытным путем было установлено, что устранение всех выявленных ранее выбросов построчно сократило бы общее число строк до недопустимого минимума.

Использование StandardScaler позволило осуществить нормализацию нецелевых признаков для снижения чувствительности моделей. При этом целевые переменные не были затронуты в процессе нормализации.

1.8 Итоги

По результатам разведывательного анализа данных были выполнены следующие действия:

1. Проведен первоначальный анализ данных, выявлены и удалены пропуски и дубликаты;
2. Построены графики распределения целевых переменных, установлен характер распределения, применено логарифмирование для улучшенной визуализации;
3. Выявлены и визуализированы выбросы в целевых переменных;
4. Осуществлен корреляционный анализ данных, проанализированы линейные зависимости между целевыми и нецелевыми переменными, сделаны предположения о значимости тех или иных признаков по результатам корреляционного анализа;
5. Выявлены выбросы в данных и удалены в целевых признаках;
6. Произведена нормализация нецелевых показателей, отобраны наиболее значимые признаки;
7. Изначальный датасет был разделен и сохранен в три набора (по целевым переменным) в формате .csv.

2 Построение регрессионных моделей

Регрессионные модели в контексте представленной задачи могут быть полезны в прогнозировании точных значений тех или иных признаков химических соединений. Однако они могут быть не настолько эффективны в определении критериев, являются ли эти соединения активными или нет. В таком случае наиболее точные показатели могут быть получены уже при использовании классификаторов.

В рамках решения задач регрессии были использованы следующие модели:

- Lasso – линейная модель с L1-регуляризацией;
- Ridge – линейная модель с L2-регуляризацией;
- ElasticNet – линейная модель с комбинированными L1 и L2-регуляризацией;
- RandomForest – случайный лес;
- SVR – метод опорных векторов для регрессии;
- XGBoost и CatBoost – градиентный бустинг.

В качестве метрик оценки качества были использованы:

- R^2 – коэффициент детерминации;
- MAE – средняя абсолютная ошибка;
- MAPE – средняя абсолютная процентная ошибка;
- MSE – средняя квадратичная ошибка.

Ключевой метрикой был принят коэффициент детерминации.

При обучении всех вышеперечисленных моделей применялась кросс-валидация.

Подбор гиперпараметров осуществлялся с использованием функции GridSearchCV.

Обучающая и тестовая выборки делились в соотношении 70 на 30.

2.1 Регрессионная модель для IC50

В рамках задачи была разработана регрессионная модель, предназначенная для прогнозирования значения параметра IC50.

По результатам обучения моделей, перечисленных в разделе 2, получены значения метрик, приведенные на рисунке 2.1 . Наиболее оптимальные значения гиперпараметров для всех моделей представлены на рисунке 2.2.

Model	MAE	MAPE	R2	MSE
Lasso	84.1698	84.9725	0.0948	13764.4197
Ridge	86.6856	70.3476	0.0594	14303.3778
ElasticNet	84.4377	89.8213	0.0884	13861.2260
Random Forest	82.7268	90.4267	0.1175	13418.9393
SVR	71.8434	28.3339	0.0649	14219.2743
XGBoost	82.2769	113.2206	0.0963	13741.7336
CatBoost	80.3522	53.7550	0.1475	12963.3327

Рисунок 2.1 – Результаты обучения моделей

```
Lasso:  {'alpha': 3}
Ridge:  {'alpha': 100}
ElasticNet:  {'alpha': 1}
Random Forest:  {'max_depth': 5, 'n_estimators': 400}
SVR:  {'C': 100, 'epsilon': 10}
XGBoost:  {'learning_rate': 0.1, 'n_estimators': 10}
CatBoost:  {'learning_rate': 0.1, 'n_estimators': 50}
```

Рисунок 2.2 – Оптимальные гиперпараметры для моделей

Полученные результаты были перенесены на график, приведенный на рисунке 2.3.

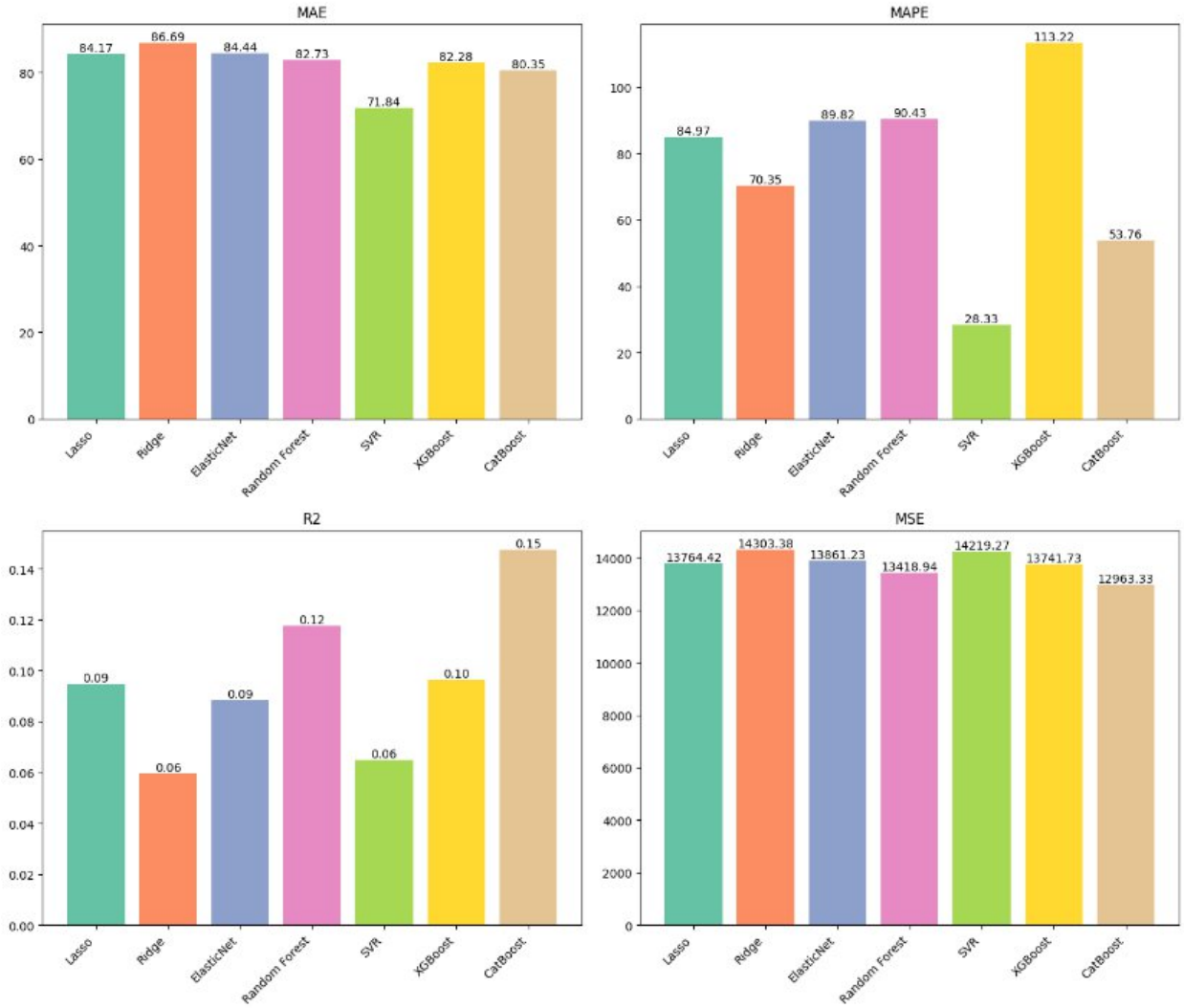

Рисунок 2.3 – Результаты обучения моделей

Как было указано в EDA, для всех трех целевых показателей характерно левоассиметричное распределение и присутствием существенного числа выбросов, устранение которых по правилу трех сигм не привело к повышению предсказательных способностей моделей. В качестве эксперимента было проведено обучение тех же моделей на логарифмированных данных, однако полученные результаты не оказались более удовлетворительными, поэтому

они не были приведены в настоящей курсовой работе. Причиной этого явления могло оказаться значительное число выбросов в наиболее ключевых для целевых показателей признаках, однако простое удаление строк, содержащих выбросы, приводит к недопустимому сокращению общего объема данных.

Из всех полученных результатов наиболее оптимальным в условиях данной задачи была выбрана модель CatBoost, продемонстрировавшая наиболее высокий коэффициент детерминации (0,15) и самый низкий MAPE.

2.2 Регрессионная модель CC50

В рамках задачи была разработана регрессионная модель, предназначенная для прогнозирования значения параметра CC50.

По результатам обучения моделей, перечисленных в разделе 2, получены значения метрик, приведенные на рисунке 2.4 . Наиболее оптимальные значения гиперпараметров для всех моделей представлены на рисунке 2.5.

Model	MAE	MAPE	R2	MSE
Lasso	308.3366	7.9118	0.3287	144203.1377
Ridge	301.1979	7.7839	0.3630	136830.1433
ElasticNet	313.8938	8.1381	0.3337	143114.3274
Random Forest	294.8103	6.5842	0.3463	140407.7880
SVR	285.6762	4.8079	0.2960	151216.3064
XGBoost	305.9449	7.2085	0.2283	165767.9935
CatBoost	299.1926	7.5954	0.3811	132942.7709

Рисунок 2.4 – Результаты обучения моделей

```
Lasso:  {'alpha': 10}
Ridge:  {'alpha': 50}
ElasticNet:  {'alpha': 1}
Random Forest:  {'max_depth': 30, 'n_estimators': 150}
SVR:  {'C': 200, 'epsilon': 10}
XGBoost:  {'learning_rate': 0.1, 'n_estimators': 25}
CatBoost:  {'learning_rate': 0.1, 'n_estimators': 50}
```

Рисунок 2.5 – Оптимальные гиперпараметры для моделей

Полученные результаты были перенесены на график, приведенный на рисунке 2.6.

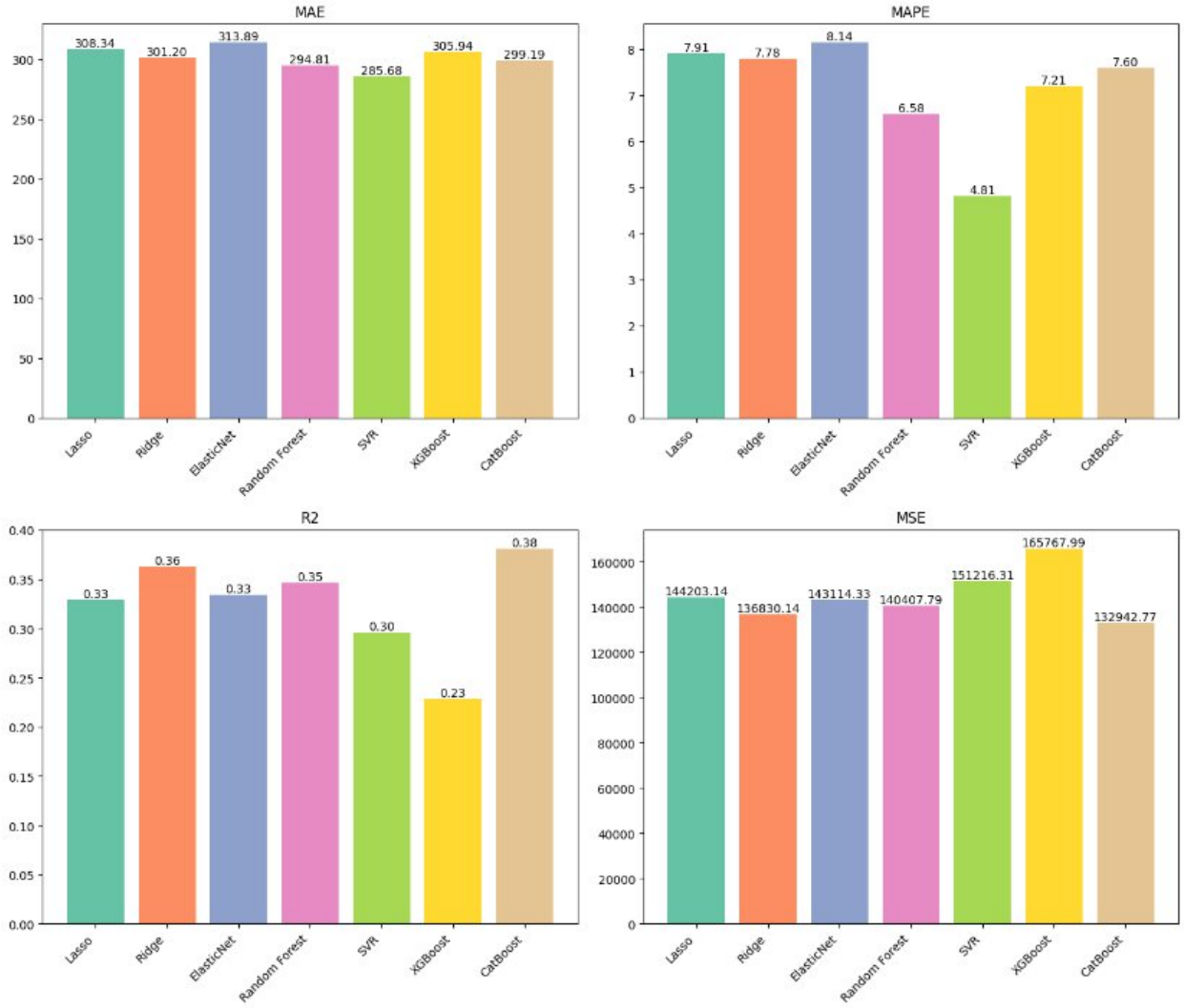

Рисунок 2.6 – Результаты обучения моделей

Как было указано в EDA, для всех трех целевых показателей характерно левоассиметричное распределение и присутствием существенного числа выбросов, устранение которых по правилу трех сигм не привело к повышению предсказательных способностей моделей. В качестве эксперимента было проведено обучение тех же моделей на логарифмированных данных, однако полученные результаты не оказались более удовлетворительными, поэтому они не были приведены в настоящей курсовой работе. Причиной этого явления могло оказаться значительное число выбросов в наиболее ключевых для целевых показателей признаках, однако простое удаление строк, содержащих выбросы, приводит к недопустимому сокращению общего объема данных.

Из всех полученных результатов наиболее оптимальным в условиях данной задачи была выбрана модель CatBoost, продемонстрировавшая наиболее высокий коэффициент детерминации (0,38).

2.3 Регрессионная модель SI

В рамках задачи была разработана регрессионная модель, предназначенная для прогнозирования значения параметра SI.

По результатам обучения моделей, перечисленных в разделе 2, получены значения метрик, приведенные на рисунке 2.7 . Наиболее

оптимальные значения гиперпараметров для всех моделей представлены на рисунке 2.8.

Model	MAE	MAPE	R2	MSE
Lasso	6.8916	2.8909	0.0247	89.2432
Ridge	6.5052	3.1265	0.0944	82.8716
ElasticNet	6.8211	2.9565	0.0479	87.1256
Random Forest	6.6354	3.0713	0.0615	85.8832
SVR	6.7797	3.2430	0.0957	82.7535
XGBoost	6.8614	3.0559	0.0026	91.2724
CatBoost	6.6499	3.0760	0.0671	85.3661

Рисунок 2.7 – Результаты обучения моделей

```
Lasso:  {'alpha': 1}
Ridge:  {'alpha': 100}
ElasticNet:  {'alpha': 1}
Random Forest:  {'max_depth': 5, 'n_estimators': 400}
SVR:  {'C': 10, 'epsilon': 5}
XGBoost:  {'learning_rate': 0.05, 'n_estimators': 10}
CatBoost:  {'learning_rate': 0.025, 'n_estimators': 100}
```

Рисунок 2.8 – Оптимальные гиперпараметры для моделей

Полученные результаты были перенесены на график, приведенный на рисунке 2.9.

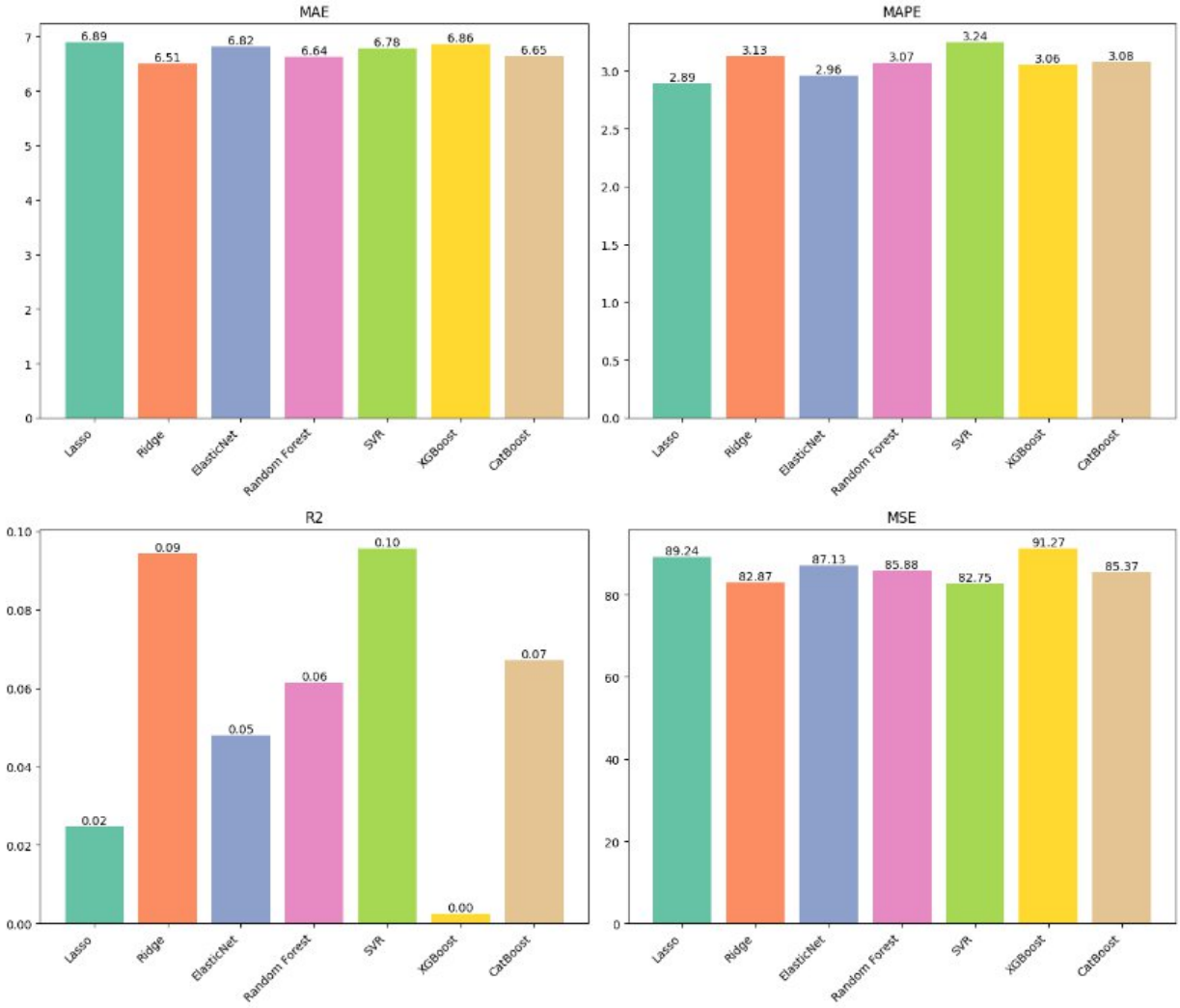

Рисунок 2.9 – Результаты обучения моделей

Как было указано в EDA, для всех трех целевых показателей характерно левоассиметричное распределение и присутствием существенного числа выбросов, устранение которых по правилу трех сигм не привело к повышению предсказательных способностей моделей. В качестве эксперимента было проведено обучение тех же моделей на логарифмированных данных, однако полученные результаты не оказались более удовлетворительными, поэтому они не были приведены в настоящей курсовой работе. Причиной этого явления могло оказаться значительное число выбросов в наиболее ключевых для целевых показателей признаках, однако простое удаление строк, содержащих выбросы, приводит к недопустимому сокращению общего объема данных.

Из всех полученных результатов наиболее оптимальным в условиях данной задачи были выбраны модели SVR и CatBoost.

3 Построение классификаторов

В рамках настоящей курсовой работы были разработаны модели классификации, способные предсказывать принадлежность каждой из целевых переменных к одному из имеющихся классов.

Решались следующие задачи классификации:

- превышает ли значение IC50 медианное значение выборки;
- превышает ли значение CC50 медианное значение выборки;
- превышает ли значение SI медианное значение выборки;
- превышает ли значение SI значение 8.

Как можно понять из определений поставленных задач, необходимо установить принадлежность каждой целевой переменной к одному из двух возможных классов. Таким образом, в данной работе требуется разработать бинарные классификаторы.

В рамках решения задач классификации были использованы следующие модели:

- LogisticRegression – логистическая регрессия;
- DecisionTree – деревья решений;
- kNearestNeighbors (kNN) – метод k-ближайших соседей;
- RandomForest – случайный лес;
- SVC – метод опорных векторов для классификации;
- XGBoost и CatBoost – градиентный бустинг.

В качестве метрик оценки качества были использованы:

- Accuracy – точность классификации;
- Precision – доля правильных положительных предсказаний среди всех предсказанных положительных;
- F1 – гармоническое среднее precision и recall;
- Recall – доля правильных положительных предсказаний среди всех реальных положительных примеров.

Ключевой метрикой была принята точность (и F1 в последней задаче).

При обучении всех вышеперечисленных моделей применялась кросс-валидация.

Обучающая и тестовая выборки делились в соотношении 70 на 30.

3.1 Классификатор IC50

В рамках задачи был разработан бинарный классификатор, предназначенный для прогнозирования принадлежности параметра IC50 к одному из двух возможных классов.

Первоначально целевой показатель был исследован на наличие дисбаланса классов. Результат исследования приведен на рисунке 3.1.

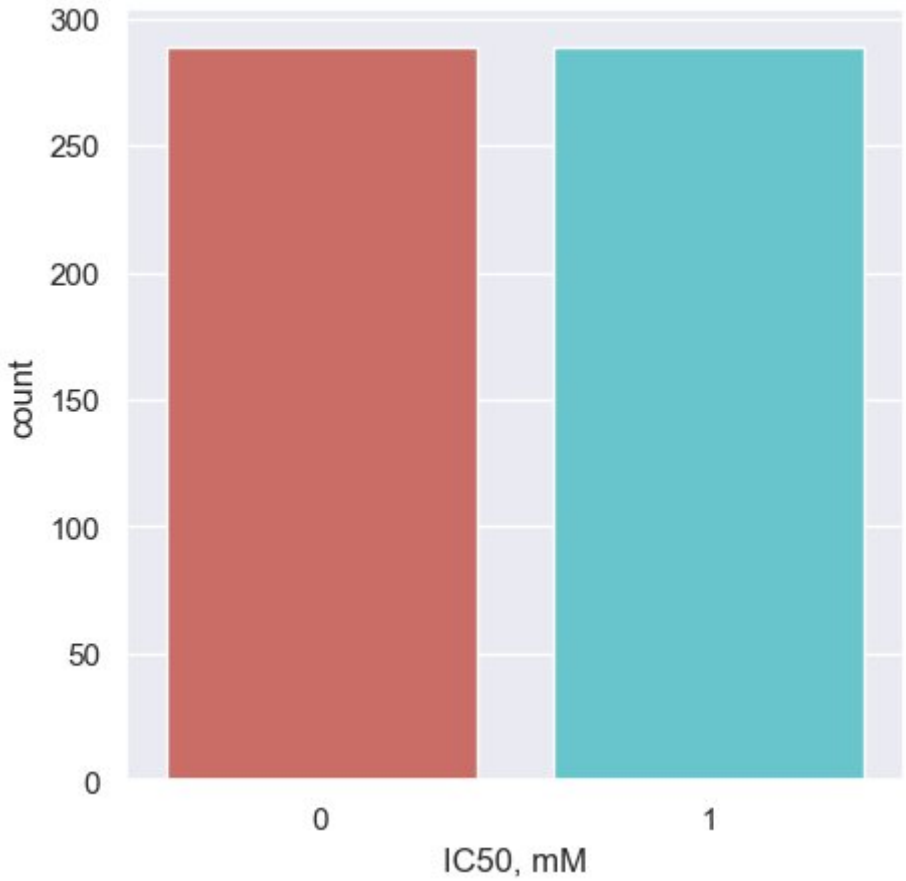

Рисунок 3.1 – Проверка целевой переменной на дисбаланс классов

Дисбаланс классов отсутствует.

По результатам обучения моделей, перечисленных в разделе 3, получены значения метрик, приведенные на рисунках 3.2 и 3.3.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
2	RandomForest	0.6471	0.6560	0.6398	0.6300	0.7137	0.6694	0.7004	0.7345
6	CatBoost	0.6610	0.6675	0.6570	0.6507	0.6895	0.6385	0.6831	0.7345
5	XGBoost	0.6610	0.6679	0.6563	0.6472	0.6815	0.6371	0.6667	0.6991
4	SVC	0.6419	0.6591	0.6217	0.5920	0.6613	0.6218	0.6379	0.6549
1	DecisionTree	0.5952	0.6092	0.5682	0.5327	0.6613	0.6179	0.6441	0.6726
0	LogisticRegression	0.6142	0.6177	0.6102	0.6057	0.6573	0.6077	0.6502	0.6991
3	KNN	0.6368	0.6420	0.6297	0.6195	0.6290	0.5814	0.6198	0.6637

Рисунок 3.2 – Результаты обучения моделей

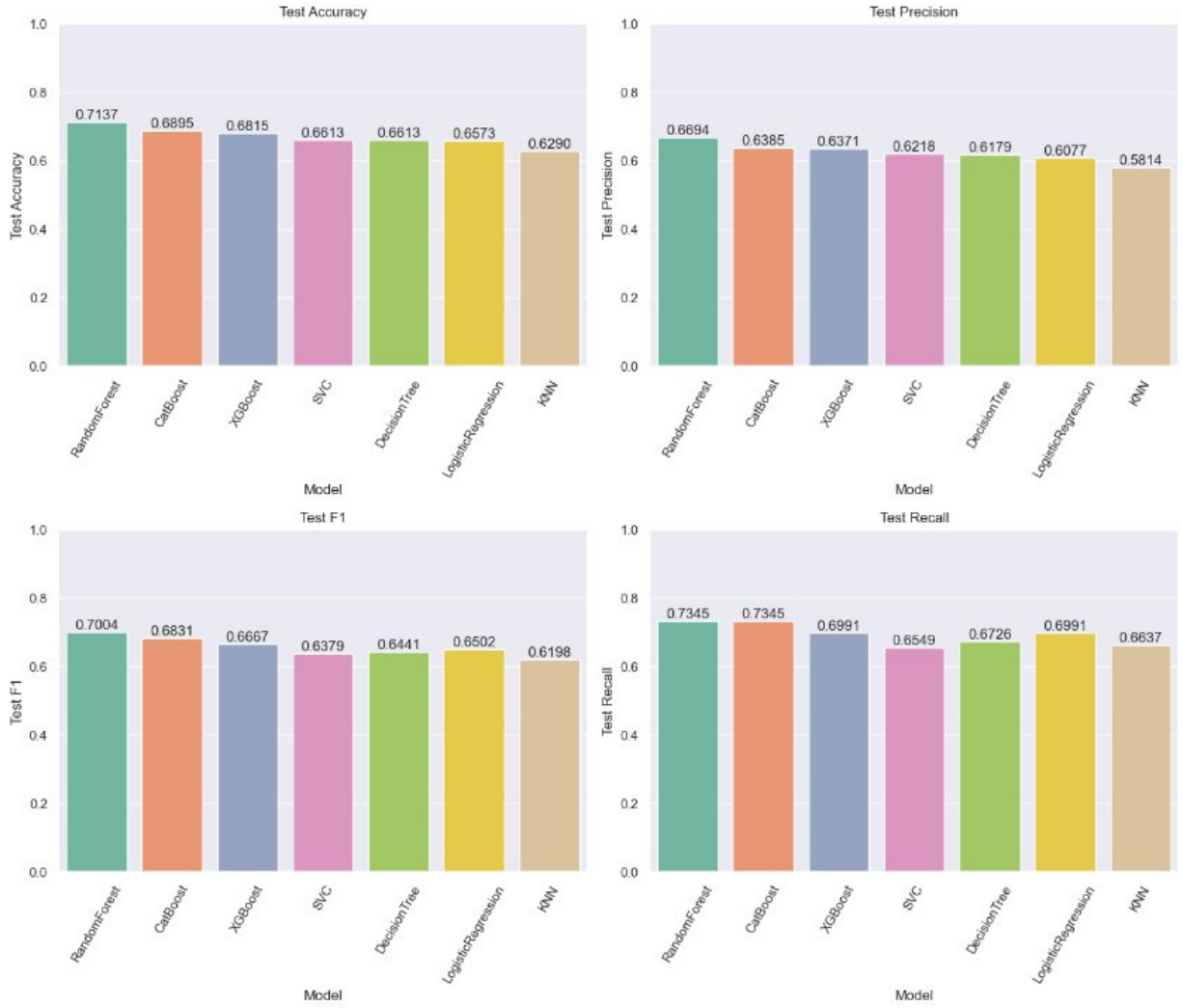

Рисунок 3.3 – Результаты обучения моделей

Модели показывают приемлемые результаты. Наиболее оптимальной из перечисленных является RandomForest.

Следует заметить, что первоначальный эксперимент не учитывал механизм подбора гиперпараметров.

С помощью функции GridSearchCV была подобрана наиболее оптимальная конфигурация гиперпараметров для всех представленных моделей, представленная на рисунке 3.4.

```
Лучшие параметры для LogisticRegression: {'C': 100, 'max_iter': 100, 'penalty': 'l2', 'solver': 'liblinear'}
Лучшие параметры для DecisionTree: {'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 2}
Лучшие параметры для RandomForest: {'max_depth': 4, 'n_estimators': 400}
Лучшие параметры для KNN: {'metric': 'manhattan', 'n_neighbors': 3}
Лучшие параметры для SVC: {'C': 10, 'gamma': 'scale', 'kernel': 'linear'}
Лучшие параметры для XGBoost: {'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 500}
Лучшие параметры для CatBoost: {'n_estimators': 500}
```

Рисунок 3.4 – Конфигурация подобранных гиперпараметров

По результатам повторного обучения моделей с подобранными гиперпараметрами получены значения метрик, приведенные на рисунках 3.5 и 3.6.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
5	XGBoost	0.6835	0.6897	0.6819	0.6783	0.7056	0.6639	0.6894	0.7168
6	CatBoost	0.6713	0.6769	0.6690	0.6645	0.6976	0.6508	0.6862	0.7257
2	RandomForest	0.6627	0.6738	0.6540	0.6403	0.6774	0.6260	0.6721	0.7257
4	SVC	0.6523	0.6624	0.6429	0.6264	0.6694	0.6220	0.6583	0.6991
3	KNN	0.6437	0.6522	0.6325	0.6160	0.6694	0.6325	0.6435	0.6549
0	LogisticRegression	0.6332	0.6335	0.6375	0.6437	0.6492	0.6000	0.6420	0.6903
1	DecisionTree	0.6593	0.6658	0.6540	0.6472	0.6452	0.6033	0.6239	0.6460

Рисунок 3.5 – Результаты обучения моделей

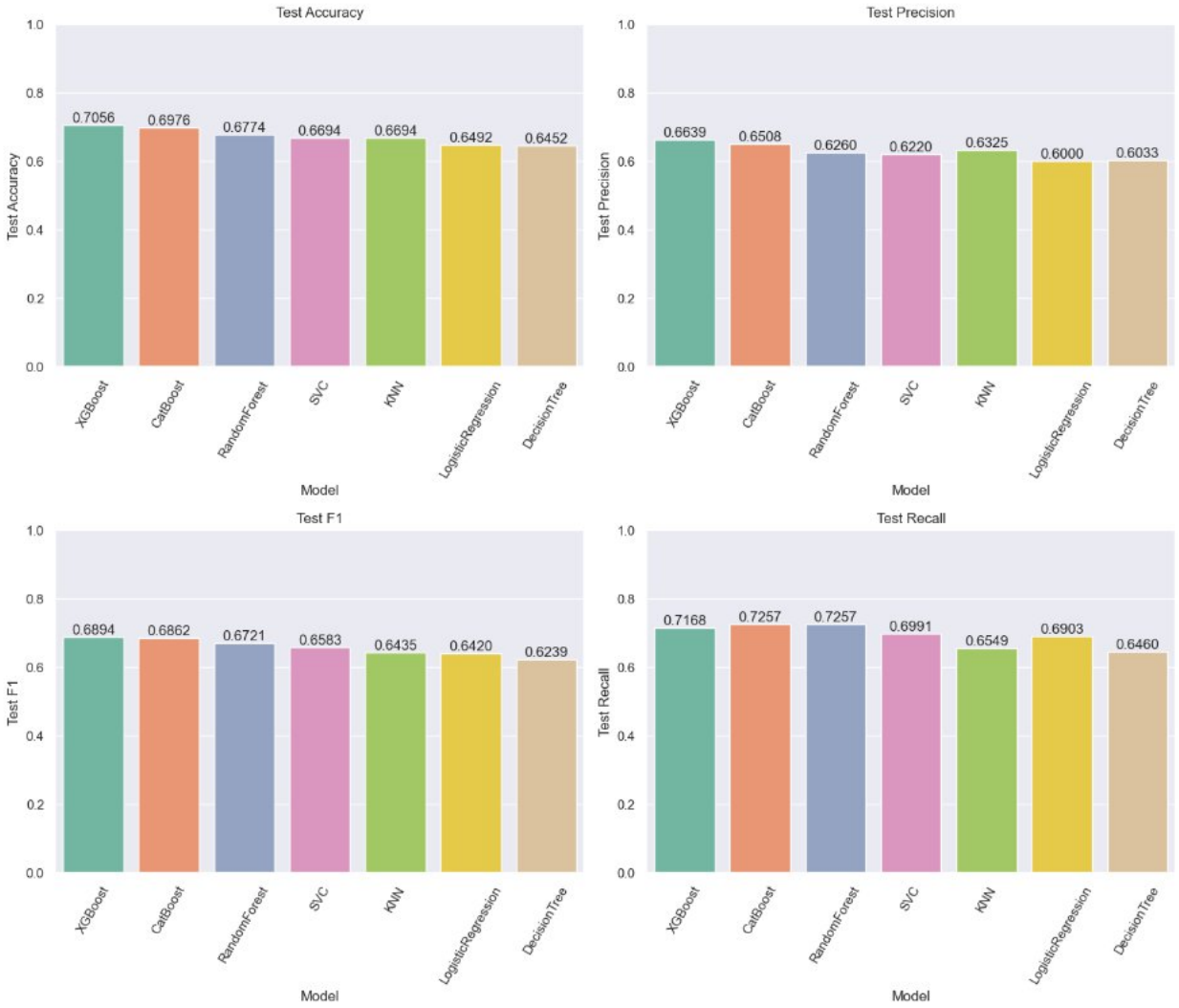

Рисунок 3.6 – Результаты обучения моделей

Модели вновь показывают приемлемые результаты. Наиболее оптимальными из перечисленных являются XGBoost и CatBoost.

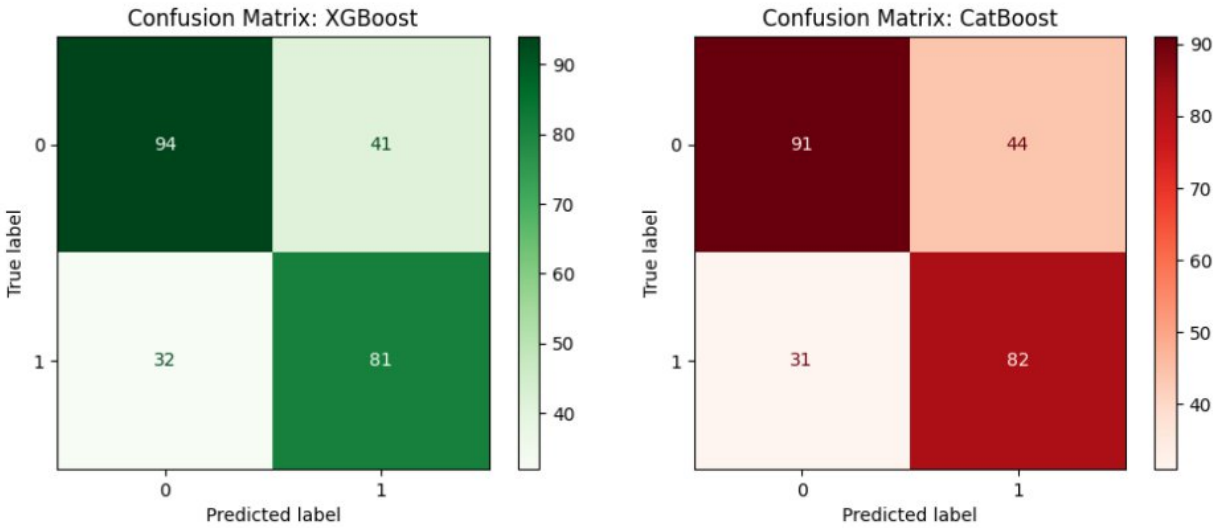

Рисунок 3.7 – Матрицы ошибок для лучших моделей

Модель XGBoost показывает наилучшую точность предсказания меток классов, а CatBoost демонстрирует лучший баланс между precision и recall.

3.2 Классификатор CC50

В рамках задачи был разработан бинарный классификатор, предназначенный для прогнозирования принадлежности параметра CC50 к одному из двух возможных классов.

Первоначально целевой показатель был исследован на наличие дисбаланса классов. Результат исследования приведен на рисунке 3.8.

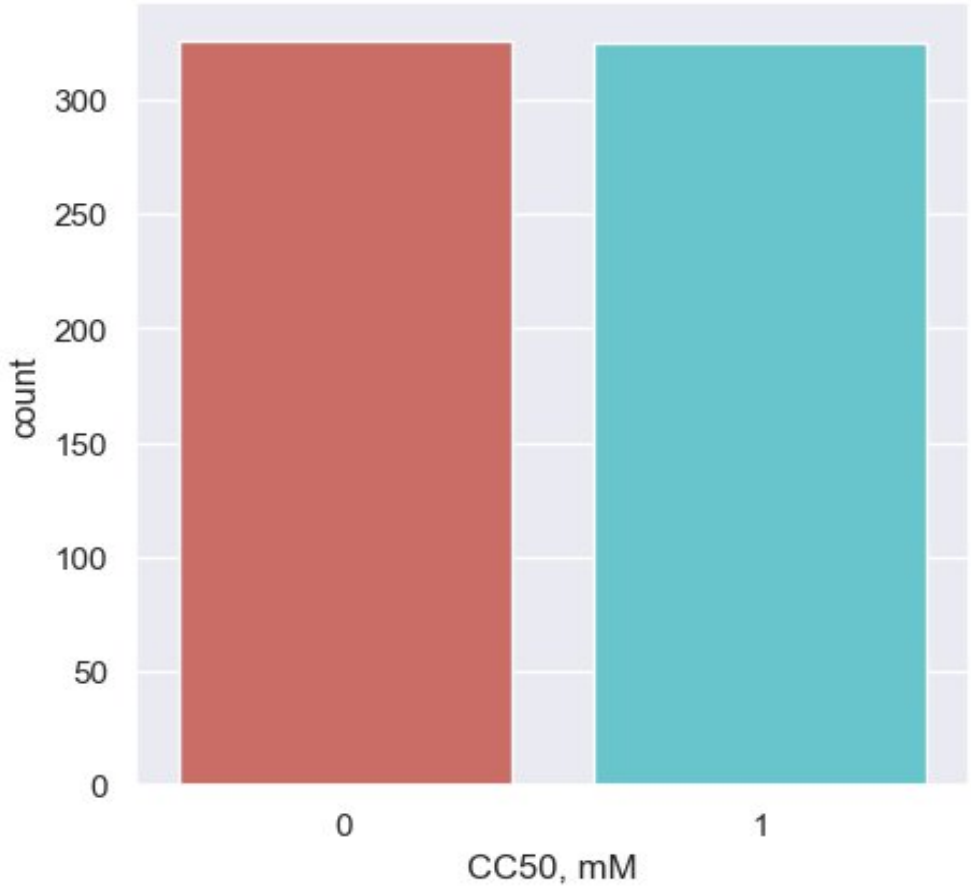

Рисунок 3.8 – Проверка целевой переменной на дисбаланс классов

Дисбаланс классов отсутствует.

По результатам обучения моделей, перечисленных в разделе 3, получены значения метрик, приведенные на рисунках 3.9 и 3.10.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
6	CatBoost	0.7235	0.7193	0.7273	0.7385	0.7714	0.7310	0.7681	0.8092
2	RandomForest	0.7281	0.7332	0.7269	0.7231	0.7679	0.7463	0.7547	0.7634
1	DecisionTree	0.6713	0.6710	0.6718	0.6738	0.7643	0.7407	0.7519	0.7634
5	XGBoost	0.7266	0.7253	0.7292	0.7354	0.7571	0.7368	0.7424	0.7481
4	SVC	0.7281	0.7207	0.7322	0.7446	0.7571	0.7059	0.7606	0.8244
0	LogisticRegression	0.7189	0.7193	0.7200	0.7231	0.7500	0.7075	0.7482	0.7939
3	KNN	0.7143	0.7132	0.7168	0.7231	0.7464	0.7239	0.7321	0.7405

Рисунок 3.9 – Результаты обучения моделей

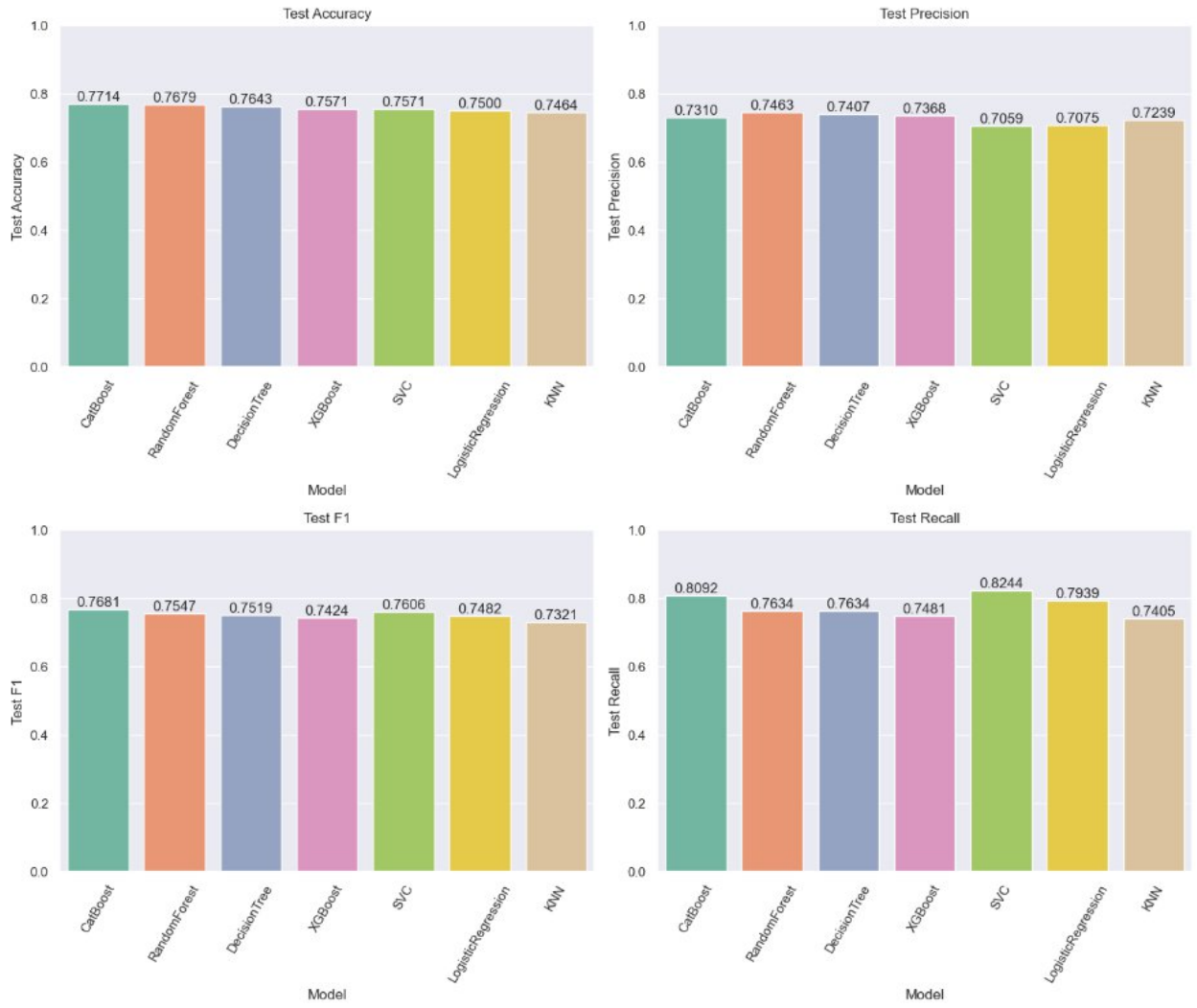

Рисунок 3.10 – Результаты обучения моделей

Модели показывают приемлемые результаты. Наиболее оптимальной из перечисленных является CatBoost.

Следует заметить, что первоначальный эксперимент не учитывал механизм подбора гиперпараметров.

С помощью функции GridSearchCV была подобрана наиболее оптимальная конфигурация гиперпараметров для всех представленных моделей, представленная на рисунке 3.11.

```
Лучшие параметры для LogisticRegression: {'C': 1, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'}
Лучшие параметры для DecisionTree: {'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 5}
Лучшие параметры для RandomForest: {'max_depth': 9, 'n_estimators': 100}
Лучшие параметры для KNN: {'metric': 'euclidean', 'n_neighbors': 3}
Лучшие параметры для SVC: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
Лучшие параметры для XGBoost: {'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 200}
Лучшие параметры для CatBoost: {'n_estimators': 100}
```

Рисунок 3.11 – Конфигурация подобранных гиперпараметров

По результатам повторного обучения моделей с подобранными гиперпараметрами получены значения метрик, приведенные на рисунках 3.12 и 3.13.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
2	RandomForest	0.7404	0.7327	0.7458	0.7600	0.7750	0.7361	0.7709	0.8092
6	CatBoost	0.7281	0.7266	0.7304	0.7354	0.7643	0.7241	0.7609	0.8015
0	LogisticRegression	0.7281	0.7264	0.7301	0.7354	0.7607	0.7162	0.7599	0.8092
1	DecisionTree	0.7189	0.7303	0.7147	0.7015	0.7571	0.7561	0.7323	0.7099
5	XGBoost	0.7389	0.7367	0.7423	0.7508	0.7571	0.7368	0.7424	0.7481
4	SVC	0.7420	0.7286	0.7496	0.7723	0.7429	0.6980	0.7429	0.7939
3	KNN	0.7250	0.7344	0.7216	0.7138	0.7000	0.6741	0.6842	0.6947

Рисунок 3.12 – Результаты обучения моделей

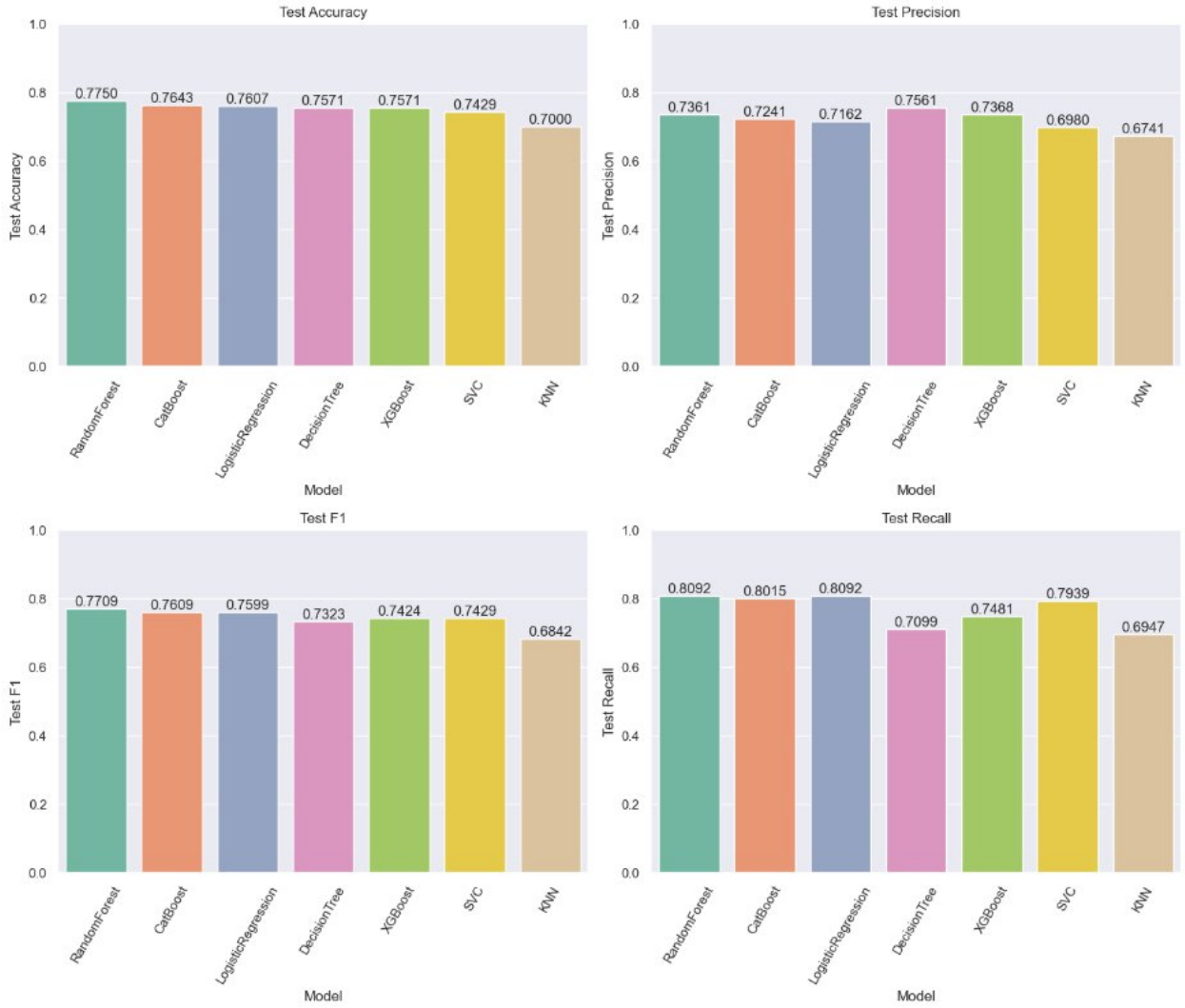

Рисунок 3.13 – Результаты обучения моделей

Модели вновь показывают приемлемые результаты. Наиболее оптимальными из перечисленных являются RandomForest и CatBoost.

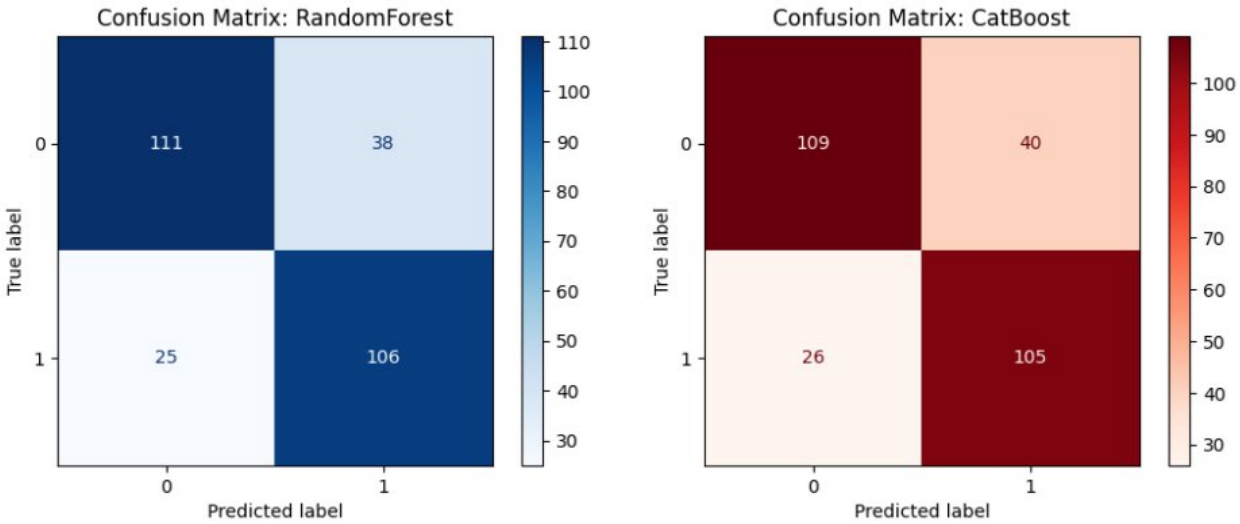

Рисунок 3.14 – Матрицы ошибок для лучших моделей

Модель RandomForest показывает наилучшую точность предсказания меток классов, а также демонстрирует лучший баланс между precision и recall.

3.3 Классификатор SI (превышение медианного значения)

В рамках задачи был разработан бинарный классификатор, предназначенный для прогнозирования принадлежности параметра SI к одному из двух возможных классов.

Первоначально целевой показатель был исследован на наличие дисбаланса классов. Результат исследования приведен на рисунке 3.15.

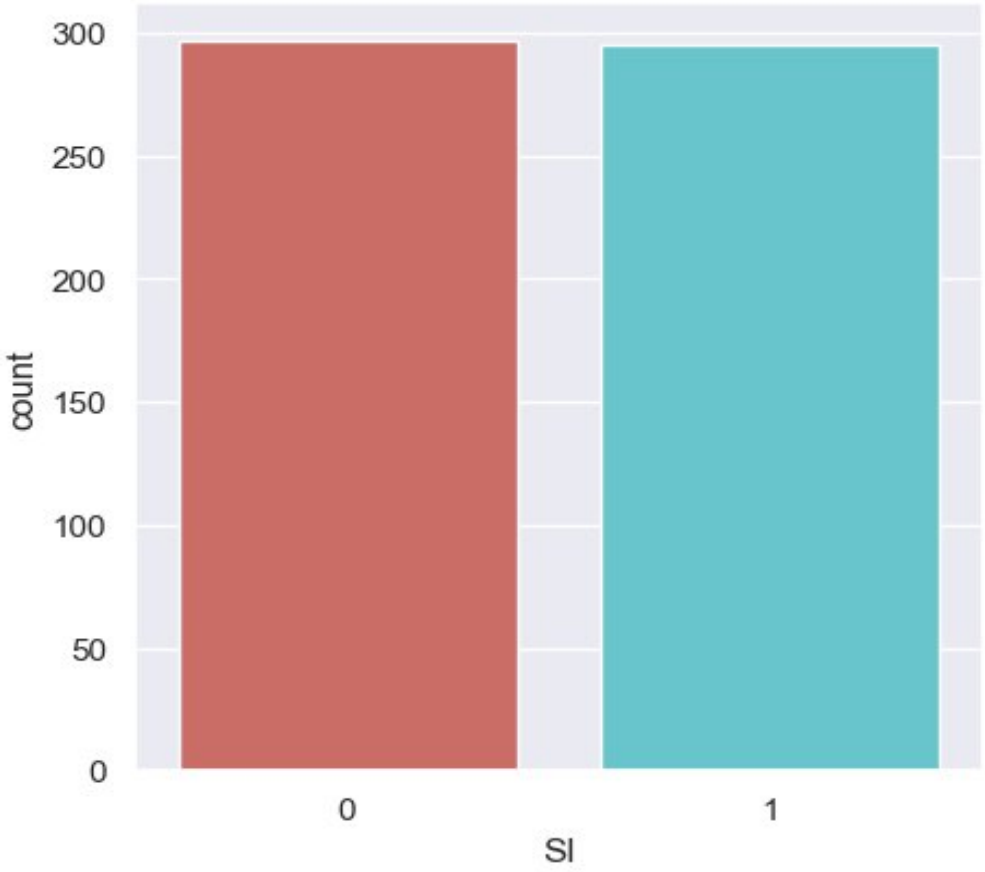

Рисунок 3.15 – Проверка целевой переменной на дисбаланс классов

Дисбаланс классов отсутствует.

По результатам обучения моделей, перечисленных в разделе 3, получены значения метрик, приведенные на рисунках 3.16 и 3.17.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
4	SVC	0.5929	0.6014	0.5708	0.5458	0.6627	0.7203	0.6641	0.6159
0	LogisticRegression	0.5591	0.5582	0.5549	0.5559	0.6235	0.6544	0.6496	0.6449
3	KNN	0.5608	0.5535	0.5821	0.6169	0.6157	0.6389	0.6525	0.6667
2	RandomForest	0.5878	0.5883	0.5778	0.5695	0.6118	0.6612	0.6178	0.5797
6	CatBoost	0.5811	0.5806	0.5725	0.5661	0.6078	0.6462	0.6269	0.6087
5	XGBoost	0.5557	0.5547	0.5473	0.5424	0.5922	0.6417	0.5969	0.5580
1	DecisionTree	0.5962	0.6073	0.5646	0.5322	0.5765	0.6415	0.5574	0.4928

Рисунок 3.16 – Результаты обучения моделей

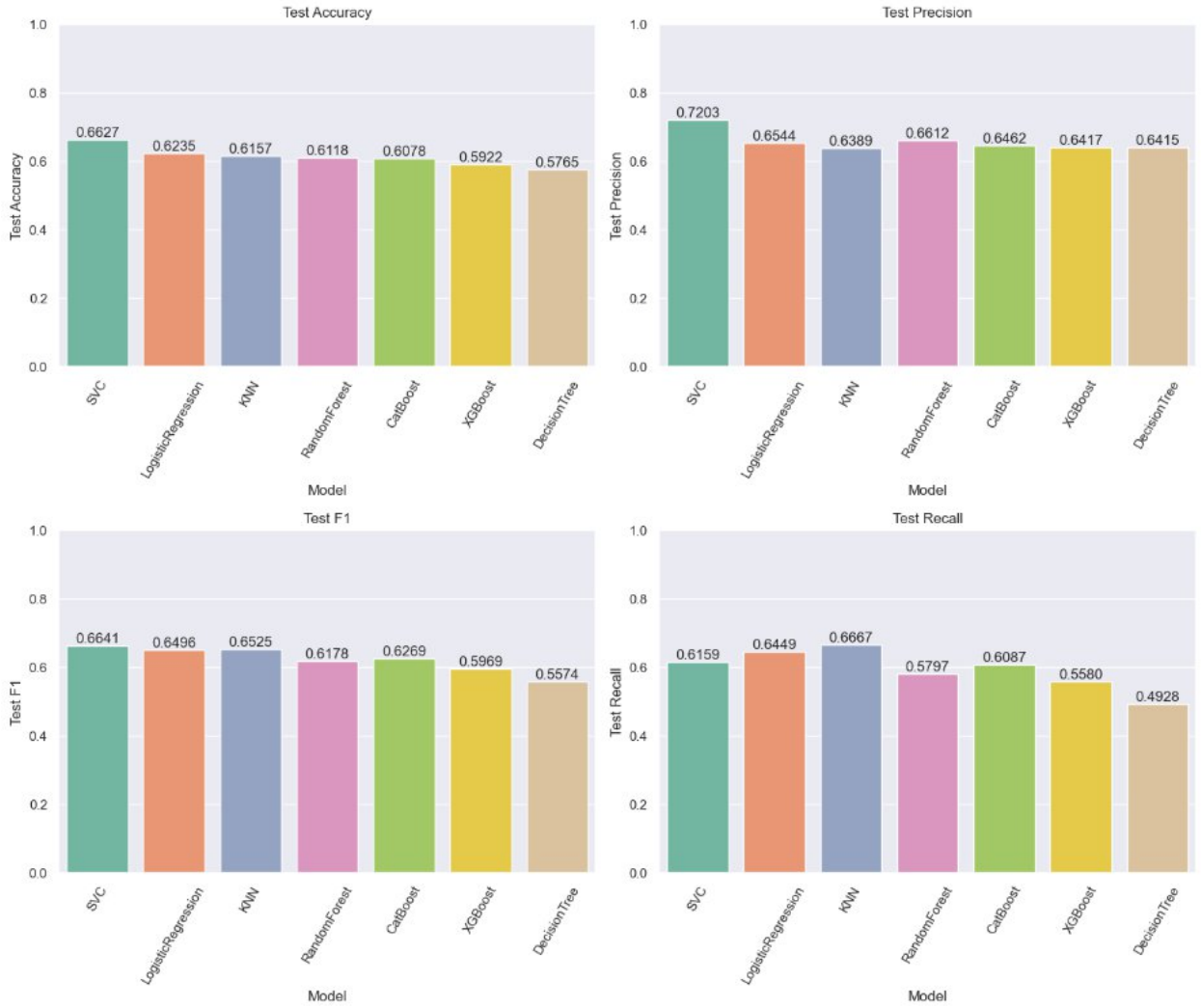

Рисунок 3.17 – Результаты обучения моделей

Модели показывают приемлемые результаты. Наиболее оптимальной из перечисленных является SVC.

Следует заметить, что первоначальный эксперимент не учитывал механизм подбора гиперпараметров.

С помощью функции GridSearchCV была подобрана наиболее оптимальная конфигурация гиперпараметров для всех представленных моделей, представленная на рисунке 3.18.

```
Лучшие параметры для LogisticRegression: {'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'liblinear'}
Лучшие параметры для DecisionTree: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2}
Лучшие параметры для RandomForest: {'max_depth': 4, 'n_estimators': 100}
Лучшие параметры для KNN: {'metric': 'manhattan', 'n_neighbors': 7}
Лучшие параметры для SVC: {'C': 0.1, 'gamma': 'scale', 'kernel': 'rbf'}
Лучшие параметры для XGBoost: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 400}
Лучшие параметры для CatBoost: {'n_estimators': 50}
```

Рисунок 3.18 – Конфигурация подобранных гиперпараметров

По результатам повторного обучения моделей с подобранными гиперпараметрами получены значения метрик, приведенные на рисунках 3.19 и 3.20.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
3	KNN	0.5844	0.5758	0.6053	0.6407	0.6471	0.6714	0.6763	0.6812
0	LogisticRegression	0.5827	0.5802	0.5816	0.5864	0.6392	0.6825	0.6515	0.6232
2	RandomForest	0.6115	0.6310	0.5779	0.5356	0.6353	0.6891	0.6381	0.5942
5	XGBoost	0.6047	0.6072	0.5954	0.5864	0.6157	0.6587	0.6288	0.6014
6	CatBoost	0.6064	0.6119	0.5907	0.5729	0.6118	0.6489	0.6320	0.6159
1	DecisionTree	0.5997	0.6173	0.5666	0.5288	0.6078	0.6759	0.5935	0.5290
4	SVC	0.6064	0.7024	0.4758	0.3627	0.5961	0.7465	0.5072	0.3841

Рисунок 3.19 – Результаты обучения моделей

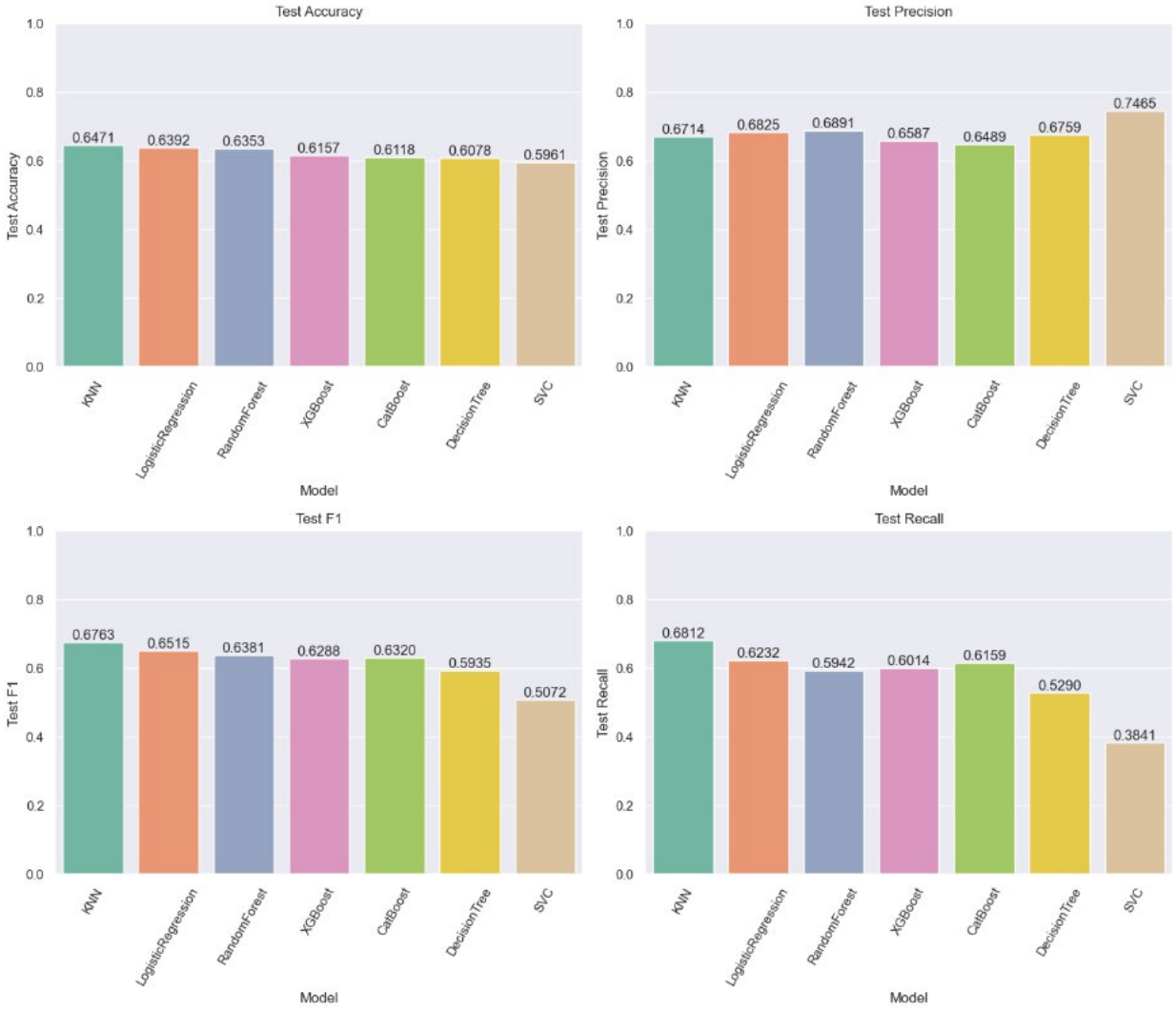

Рисунок 3.20 – Результаты обучения моделей

Модели вновь показывают приемлемые результаты. Наиболее оптимальной из перечисленных является kNN.

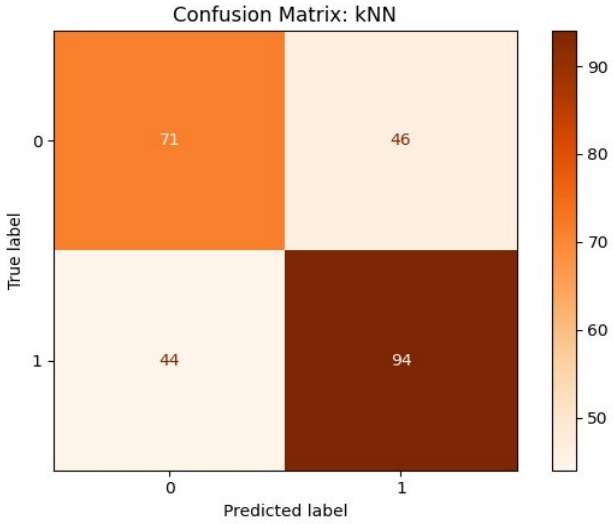

Рисунок 3.21 – Матрица ошибок для лучшей модели

Модель kNN показывает наилучшую точность предсказания меток классов, а также демонстрирует лучший баланс между precision и recall.

3.4 Классификатор SI (превышение значения 8)

В рамках задачи был разработан бинарный классификатор, предназначенный для прогнозирования принадлежности параметра SI к одному из двух возможных классов.

Первоначально целевой показатель был исследован на наличие дисбаланса классов. Результат исследования приведен на рисунке 3.22.

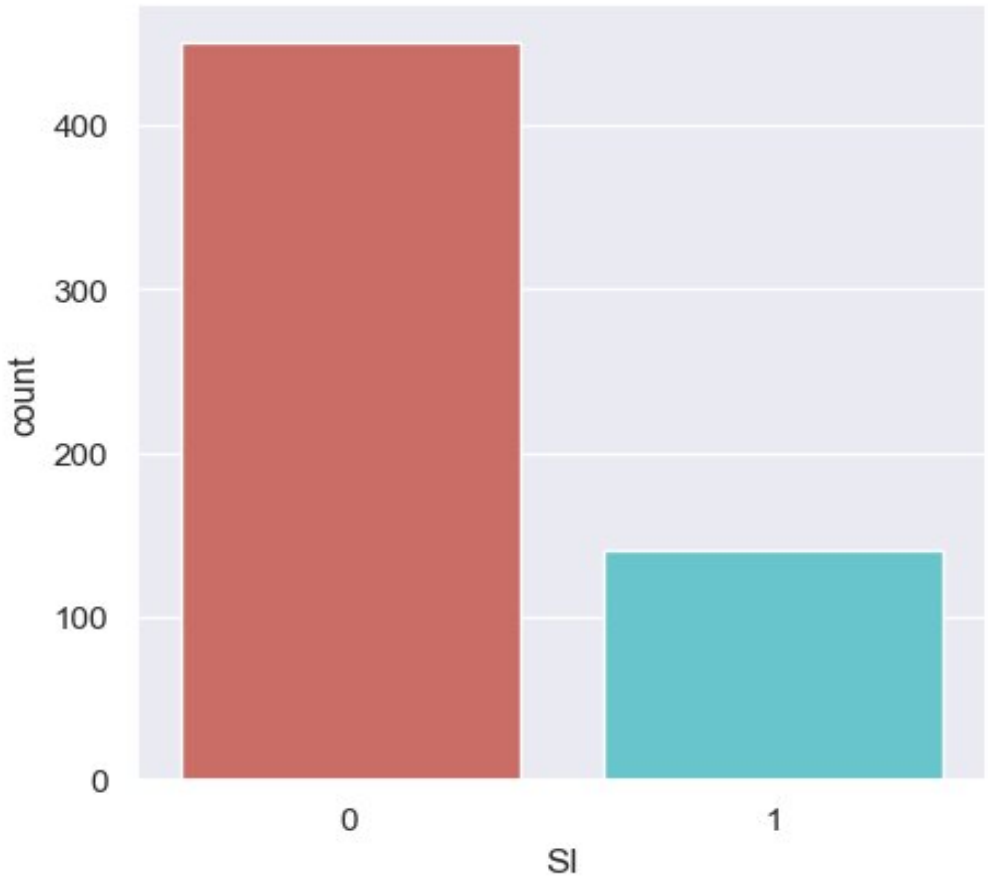

Рисунок 3.22 – Проверка целевой переменной на дисбаланс классов

Присутствует заметный дисбаланс классов. Несбалансированные данные понижают предсказательную способность моделей, с большей вероятностью наши модели будут предсказывать значение 0.

Для устранения дисбаланса был использован алгоритм синтетического генерирования данных (SMOTE), при котором создавались дополнительные выборки на основе миноритарного класса. По результатам использования SMOTE и повторного разделения данных на обучающую и тестовую выборки получилось следующее распределение целевой переменной, представленной на рисунках 3.23 и 3.24.

```
Features shape after SMOTE: (1248, 145)
Classes distribution after SMOTE:
SI
0     624
1     624
Name: count, dtype: int64
Train dataset size: (873, 145), (873, 1)
Train dataset size: (375, 145), (375, 1)
```

Рисунок 3.23 – Результат работы SMOTE

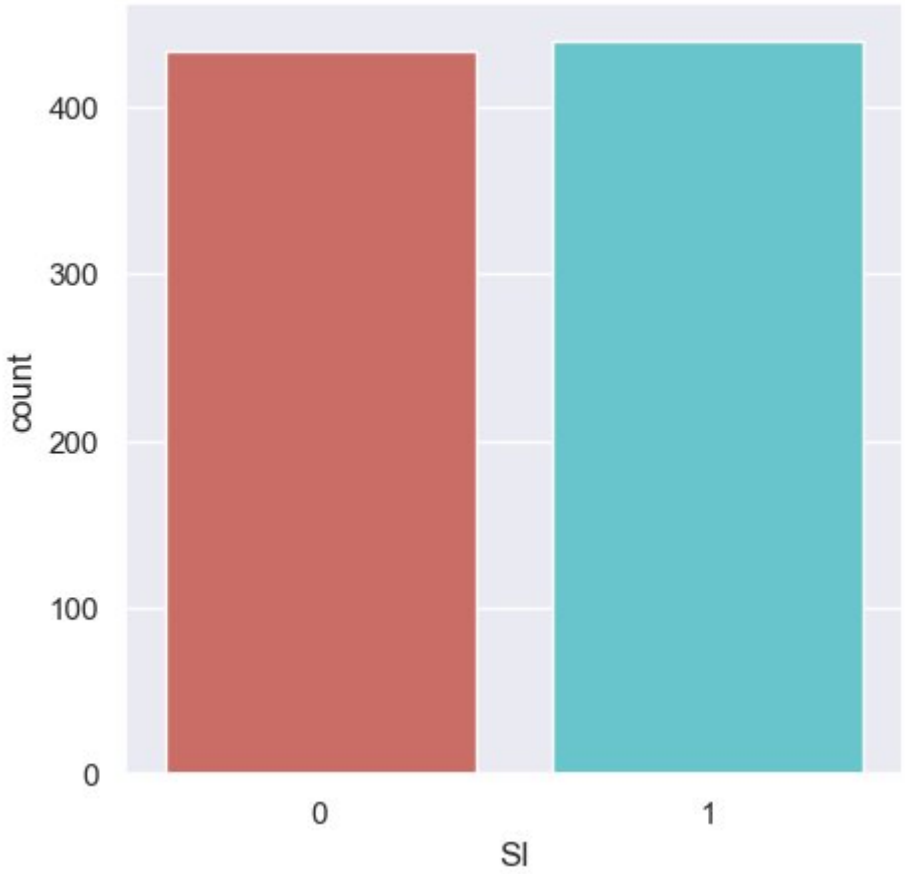

Рисунок 3.24 – Распределение целевой переменной после применения SMOTE

По результатам обучения моделей, перечисленных в разделе 3, получены значения метрик, приведенные на рисунках 3.25 и 3.26.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
5	XGBoost	0.7789	0.7696	0.7851	0.8023	0.8213	0.8063	0.8213	0.8370
6	CatBoost	0.7744	0.7625	0.7805	0.8000	0.8133	0.7938	0.8148	0.8370
2	RandomForest	0.7835	0.7796	0.7857	0.7932	0.7920	0.7880	0.7880	0.7880
4	SVC	0.6930	0.6785	0.7071	0.7409	0.7600	0.7350	0.7656	0.7989
1	DecisionTree	0.7331	0.7247	0.7411	0.7591	0.7280	0.7113	0.7302	0.7500
3	KNN	0.6930	0.6557	0.7286	0.8205	0.6827	0.6419	0.7119	0.7989
0	LogisticRegression	0.6598	0.6515	0.6706	0.6932	0.6747	0.6505	0.6872	0.7283

Рисунок 3.25 – Результаты обучения моделей

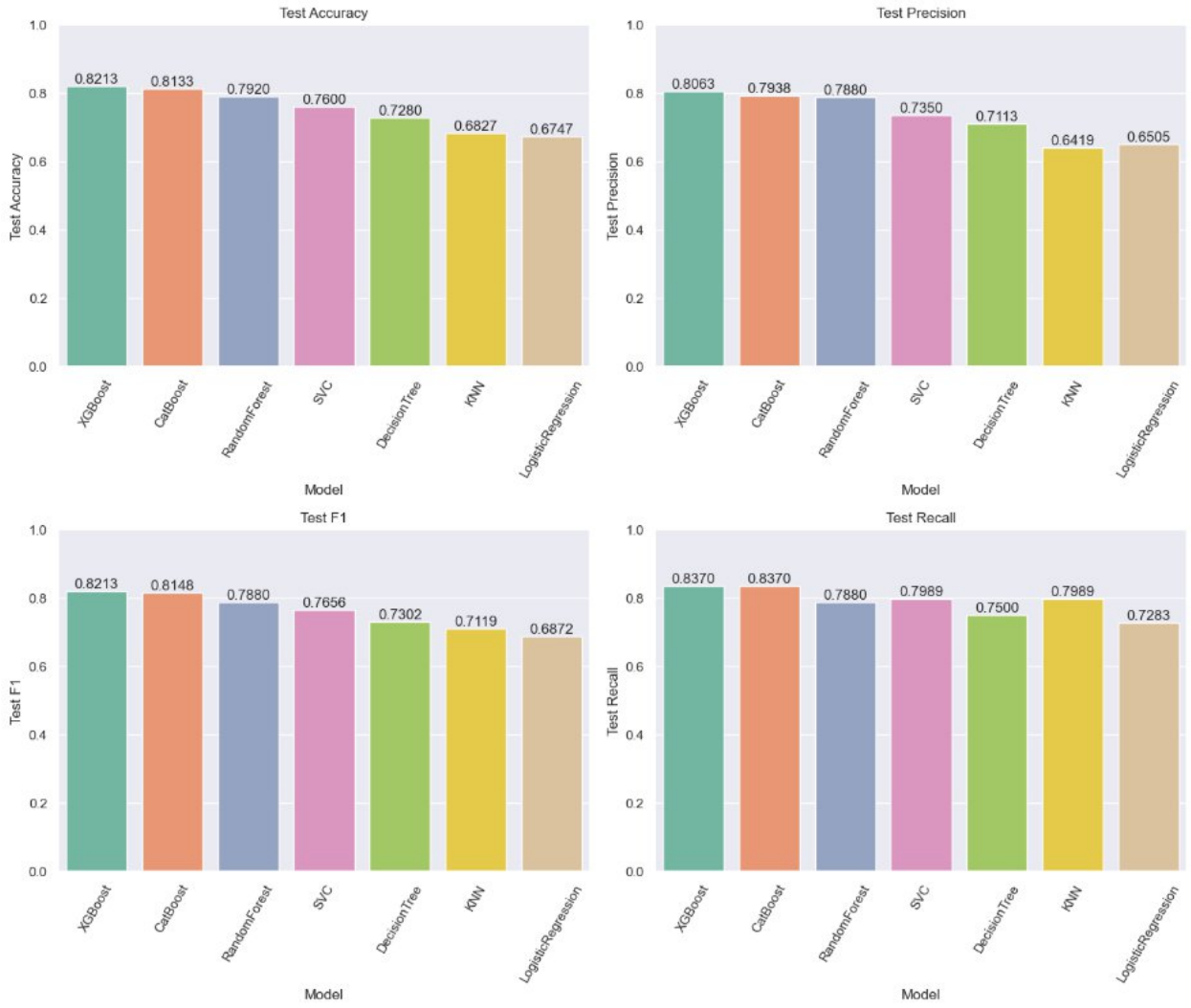

Рисунок 3.26 – Результаты обучения моделей

Модели показывают приемлемые результаты. Наиболее оптимальной из перечисленных является XGBoost.

Следует заметить, что первоначальный эксперимент не учитывал механизм подбора гиперпараметров.

С помощью функции GridSearchCV была подобрана наиболее оптимальная конфигурация гиперпараметров для всех представленных моделей.

По результатам повторного обучения моделей с подобранными гиперпараметрами получены значения метрик, приведенные на рисунках 3.27 и 3.28.

	Model	CV Accuracy	CV Precision	CV F1	CV Recall	Test Accuracy	Test Precision	Test F1	Test Recall
6	CatBoost	0.7859	0.7741	0.7918	0.8114	0.8213	0.8000	0.8232	0.8478
5	XGBoost	0.7904	0.7810	0.7957	0.8114	0.8053	0.7846	0.8074	0.8315
2	RandomForest	0.7927	0.7834	0.7968	0.8114	0.8000	0.7946	0.7967	0.7989
4	SVC	0.7789	0.7386	0.7975	0.8682	0.7707	0.7311	0.7828	0.8424
1	DecisionTree	0.7400	0.7322	0.7467	0.7636	0.7280	0.7113	0.7302	0.7500
3	KNN	0.7068	0.6729	0.7354	0.8114	0.6933	0.6468	0.7255	0.8261
0	LogisticRegression	0.6827	0.6734	0.6945	0.7182	0.6800	0.6569	0.6907	0.7283

Рисунок 3.27 – Результаты обучения моделей

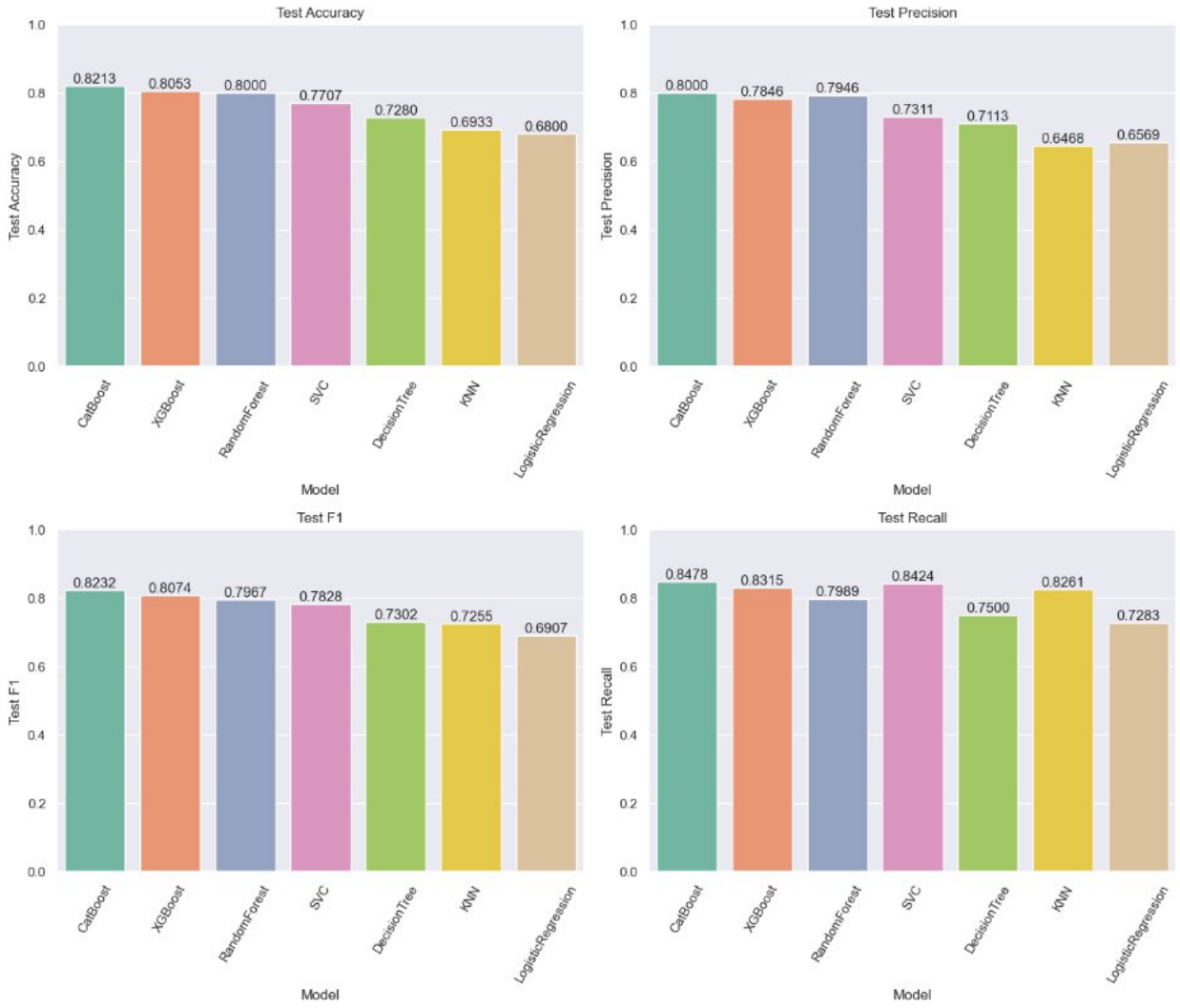

Рисунок 3.28 – Результаты обучения моделей

Модели вновь показывают приемлемые результаты. Наиболее оптимальными из перечисленных являются CatBoost и XGBoost.

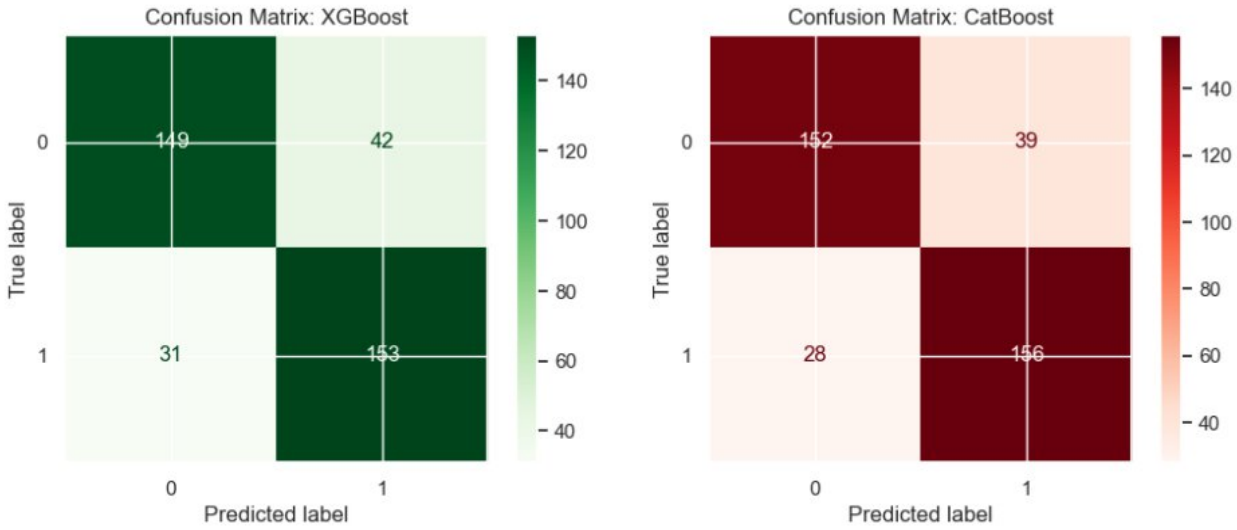

Рисунок 3.29 – Матрицы ошибок для лучших моделей

Модель CatBoost показывает наилучшую точность предсказания меток классов, а также демонстрирует лучший баланс между precision и recall.

Заключение

В рамках настоящей курсовой работы были решены поставленные задачи по созданию моделей машинного обучения для прогнозирования активности, токсичности и селективности химических соединений. Перед созданием моделей был выполнен разведывательный анализ данных.

По результатам разработки регрессионных моделей были получены следующие показатели:

Таблица 2 – Результаты разработки регрессионных моделей

Задача	Выбранная модель	R^2
IC_{50}	CatBoost	0.15
CC_{50}	CatBoost	0.38
SI	SVR	0.10

Низкие значения коэффициентов детерминации указывают на обилие выбросов в данных и недостаточного объема выборки.

По результатам разработки бинарных классификаторов были получены следующие показатели:

Таблица 3 – Результаты разработки классификаторов

Задача	Выбранная модель	Accuracy	Recall
IC_{50} > медиана	XGBoost	0.7056	0.7168
CC_{50} > медиана	Random Forest	0.7750	0.8092
SI > медиана	kNN	0.6471	0.6812
SI > 8	CatBoost	0.8232 (F1)	0.8478

Результаты работы классификаторов во всех задачах следует признать приемлемыми и подходящими для прогнозирования принадлежности целевых показателей к тем или иным меткам классов.

В качестве рекомендаций по дальнейшему повышению качества работы моделей может служить увеличение общего объема первоначальной выборки. Помимо этого, предсказательную способность может повысить более глубокая проработка нецелевых признаков на этапе EDA и преобразование выбросов, а также генерация полиноминальных признаков.

В конечном итоге, разработанный подход к решению задачи является перспективным и способен упорядочивать химические соединения по рассмотренным критериям. Бинарные классификаторы продемонстрировали свою эффективность в этих аспектах, регрессоры же требует дополнительных проработок для повышения точности предсказания.