# Многокритериальный поиск эффективных управленческих решений для повышения выживаемости и ускорения роста компаний в изменяющихся рыночных условиях методами анализа данных и машинного обучения


Выполнил — Минина А.В
Студент гр. 5040103/00301

Руководитель
к.ф-м.н, доцент ВШ Искусственного Интеллекта СПБПУ — Лукашин А.А.

Консультант — Dr. Sebastian Kortmann, University van Amsterdam

# Цель работы

Методами машинного обучения создать модель, позволяющую выделять наиболее ценные атрибуты управленческих решений, приводящих к повышению выживаемости и ускорению роста компаний в условиях изменяющихся внешних факторов на основе анализа данных о деятельности компаний и альянсов, исследовать работу модели на реальных данных.

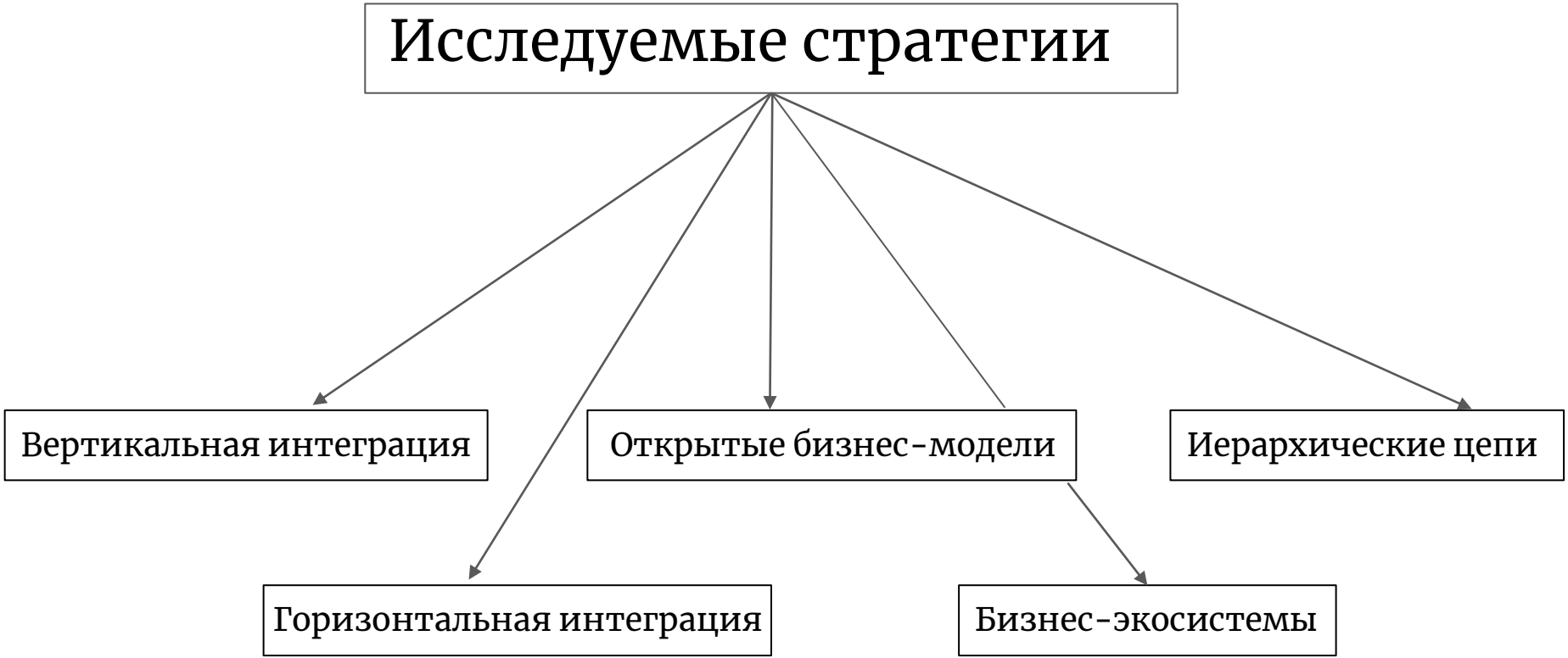
# Решаемые в проекте задачи

- Анализ существующих экономических бизнес-стратегий;
- Анализ исходных данных, исследование влияния независимых внешних переменных на характеристики выживаемости и роста компаний и альянсов, выявление наиболее важных факторов;
- Построение модели множественной линейной регрессии для оценки независимых параметров и их взаимодействий;
- Построение модели для алгоритма машинного обучения, осуществляющего выявление наиболее эффективных деревьев решений, используя соотношение данных по деятельности отдельных компаний и выживаемости альянсов;
- Программная реализация алгоритма машинного обучения;
- Исследование работы алгоритма на реальных данных и сравнение полученных результатов.

# Исходные данные для анализа

- Источники данных
  - Данные из Bureau van Dijk, которые отражают характеристики деятельности компаний на мировом рынке;
  - Данные по альянсам, которые отражают совокупные характеристики альянсов из нескольких компаний.

- Методы и подходы к решению поставленной задачи
  - Обработка больших массивов данных;
  - Учет относительного влияния множества независимых внешних переменных;
  - Применение методов машинного обучения для нахождения наилучших рабочих параметров модели.


Исследуемые стратегии
Вертикальная интеграция
Открытые бизнес-модели
Иерархические цепи
Горизонтальная интеграция
Бизнес-экосистемы

# Предварительный анализ исходных данных

| Alliance_Deal_Name | Allian… | Bu… | Act… | Par… | Primary_SIC_… | Nation_of_Allia… | Participant_Nation |
|---|---|---|---|---|---|---|---|
| JOHN WOOD GROUP PLC/… | 1990… | Pv… | Re… | Joh… | 7699 | United Kingdom | United Kingdom… |
| SCOTT/SANYO-KOKUSAKU… | 1990… | Ma… | Lic… | Sc… | 2844 | Japan | United States J… |
| INTELLICALL/MESSAGEPH… | 1990… | Ma… | Lic… | Int… | 3661 | United States | United States U… |
| NIHON UNISYS LTD. {MITS… | 1990… | Do… | Ma… | Mit… | 3572 | Japan | Japan United St… |
| HITACHI BORDEN CHEMIC… | 1990… | Mn… | Ma… | Hit… | 3081 | Japan | Japan United St… |
| CONTINENTAL/AIR MICRO… | 1990… | Pa… | | Co… | 4512 | United States | United States M… |
| SSANGYONG PAPER CO. | 1990… | Do… | Ma… | Ss… | 2611 | South Korea | South Korea Un… |
| PT INDONESIA ASAHAN AL… | 1990… | Nat… | Ma… | Ind… | 3365 | Indonesia | Indonesia Japa… |
| SEDONA/NADO ELECTRO… | 1990… | Mn… | Lic… | Se… | 6794 | Supranational | United States S… |
| ALLIED-LYONS/GREENALL … | 1990… | Pro… | Ma… | Gr… | 5181 | | United Kingdom… |
| BRITISH PETROLEUM/EXX… | 1990… | Int… | Ma… | Brit… | 2911 | United States | United Kingdom… |
| PHYSICAL OPTICS/CHUGA… | 1990… | Co… | Ma… | Ph… | | | United States J… |
| WESTEL RADIO TELEPHO… | 1990… | Pro… | | US… | 3663 | Hungary | United States H… |
| GREEN CROSS/XOMA-STR… | 1990… | Pv… | Lic… | Xo… | 6794 | Supranational | United States J… |
| QUIDEL/TAKEDA CHEMICA… | 1990… | Mn… | Ma… | Mo… | 8742 | Japan | United States J… |
| ANIXTER-ROTELCOM, INC.… | 1990… | Wh… | Ma… | Ani… | 5065 | United States | United States U… |
| GALACTIC RESOURCES/P… | 1990… | Gol… | Ex… | Gal… | 1041 | United States | Canada Canada |

Рис. 1 - пример данных по альянсам

Рис. 2 - пример данных по Bureau van Dijk

| Alliance_Deal_Name | FIXED ASSETS (fias) | INTANGI… | TANGIB… | OTHER FI… | CURRENT … | STOC… | OTHER CURRENT A… |
|---|---|---|---|---|---|---|---|
| ADOBE SYSTEMS/C… | 13911643000.0 | 12650049… | 1075072… | 18652200… | 4857039000.0 | 0.0 | 3541461000.0 |
| ADOBE SYSTEMS/C… | 42673.0 | 0.0 | 0.0 | 42673.0 | 10070852.0 | 589075… | 4180098.0 |
| ADOBE SYSTEMS I… | 13911643000.0 | 12650049… | 1075072… | 18652200… | 4857039000.0 | 0.0 | 3541461000.0 |
| ADOBE SYSTEMS I… | 35896098319.0 | 31085831… | 2552039… | 22582271… | 109256912… | 0.0 | 5762121444.0 |
| ACCENTURE PLC /G… | | | | | | | |
| ACCENTURE PLC /G… | | | | | | | |
| SUN MICROSYSTE… | 10247000000.0 | 23110000… | 2697000… | 52390000… | 7934000000.0 | 104900… | 3930000000.0 |
| SUN MICROSYSTE… | 409800000.0 | 32040000… | 6690000… | 22500000.0 | 841200000.0 | 203300… | 379100000.0 |
| ACCENTURE LTD/S… | 10247000000.0 | 23110000… | 2697000… | 52390000… | 7934000000.0 | 104900… | 3930000000.0 |
| ACCENTURE LTD/S… | | | | | | | |
| SYMANTEC CORP/I… | 952345.0 | 0.0 | 952345.0 | 0.0 | 542296.0 | 28905.0 | 513391.0 |
| SYMANTEC CORP/I… | 92789451.0 | 21649792.0 | 7113965… | 0.0 | 1229005516.0 | 758131… | 1097894638.0 |
| SOFTBANK CORP/A… | | | | | | | |
| SOFTBANK CORP/A… | 639780000.0 | 54447800… | 3280600… | 62496000.0 | 273361000.0 | 0.0 | 241105000.0 |
| PROGRESS SOFTW… | 378482.0 | 3693.0 | 374789.0 | 0.0 | 12385445.0 | 0.0 | 7244602.0 |
| PROGRESS SOFTW… | | | | | | | |
| ABB LTD/DASSAULT… | 762440641.0 | 57483931.0 | 2341049.0 | 70261566… | 58851236.0 | 0.0 | 26975388.0 |
| ABB LTD/DASSAULT… | 40462705.0 | 20508824.0 | 1995388… | 0.0 | 121846895.0 | 402603… | 46909652.0 |
| DASSAULT SYSTEM… | 762440641.0 | 57483931.0 | 2341049.0 | 70261566… | 58851236.0 | 0.0 | 26975388.0 |
| DASSAULT SYSTEM… | 77326539671.0 | 67370654… | 6789972… | 87543787… | 233247100… | 383162… | 17093876138.0 |
| VISA INC/AUTOMATI… | 10459673.0 | 993597.0 | 9466076.0 | 0.0 | 133091805.0 | 349732.0 | 114141719.0 |

# Применяемый расчет параметров SIC-кодов

1. Для построения треугольников из кодов был выбран метод Евклидова расстояния. Для вычисления координат используется остаток от деления и целая часть от деления. Стандартная формула Евклидова расстояния выглядит так:

$$distance(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

2. Вычисляем периметры треугольников по следующей формуле:

$$P = \frac{1}{2}A + B + C$$

3. Находим площади треугольников по формуле Герона:

$$S = \sqrt{P(P - A)(P - B)(P - C)}$$

# Обработка естественного языка

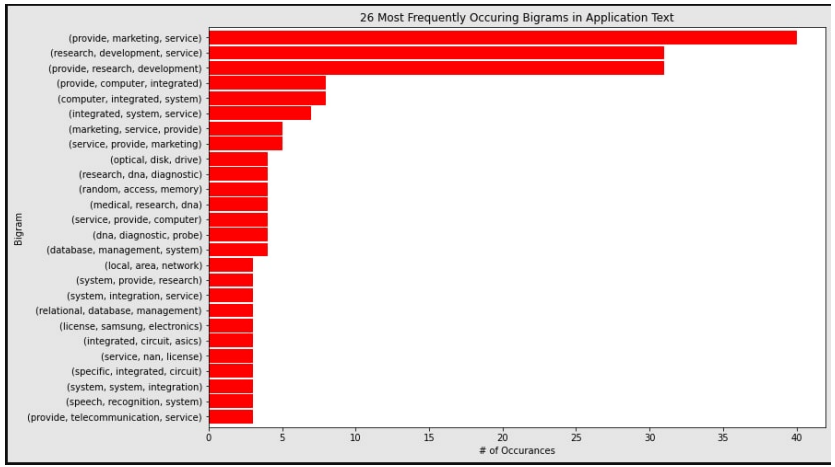
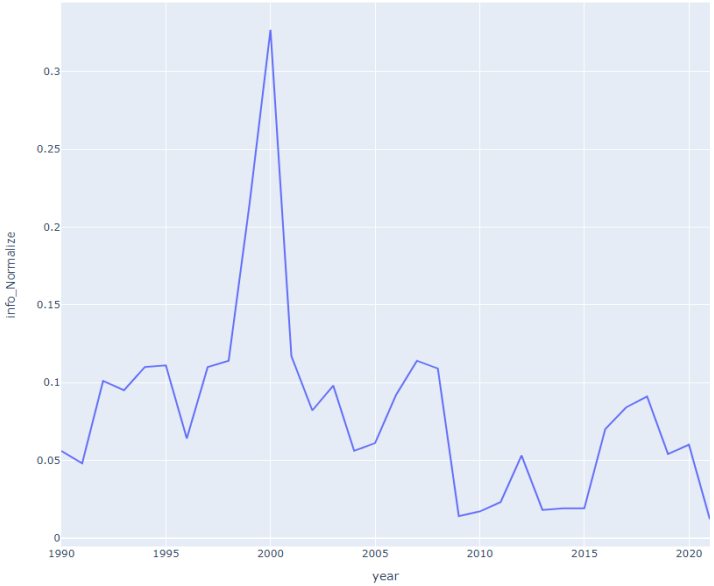
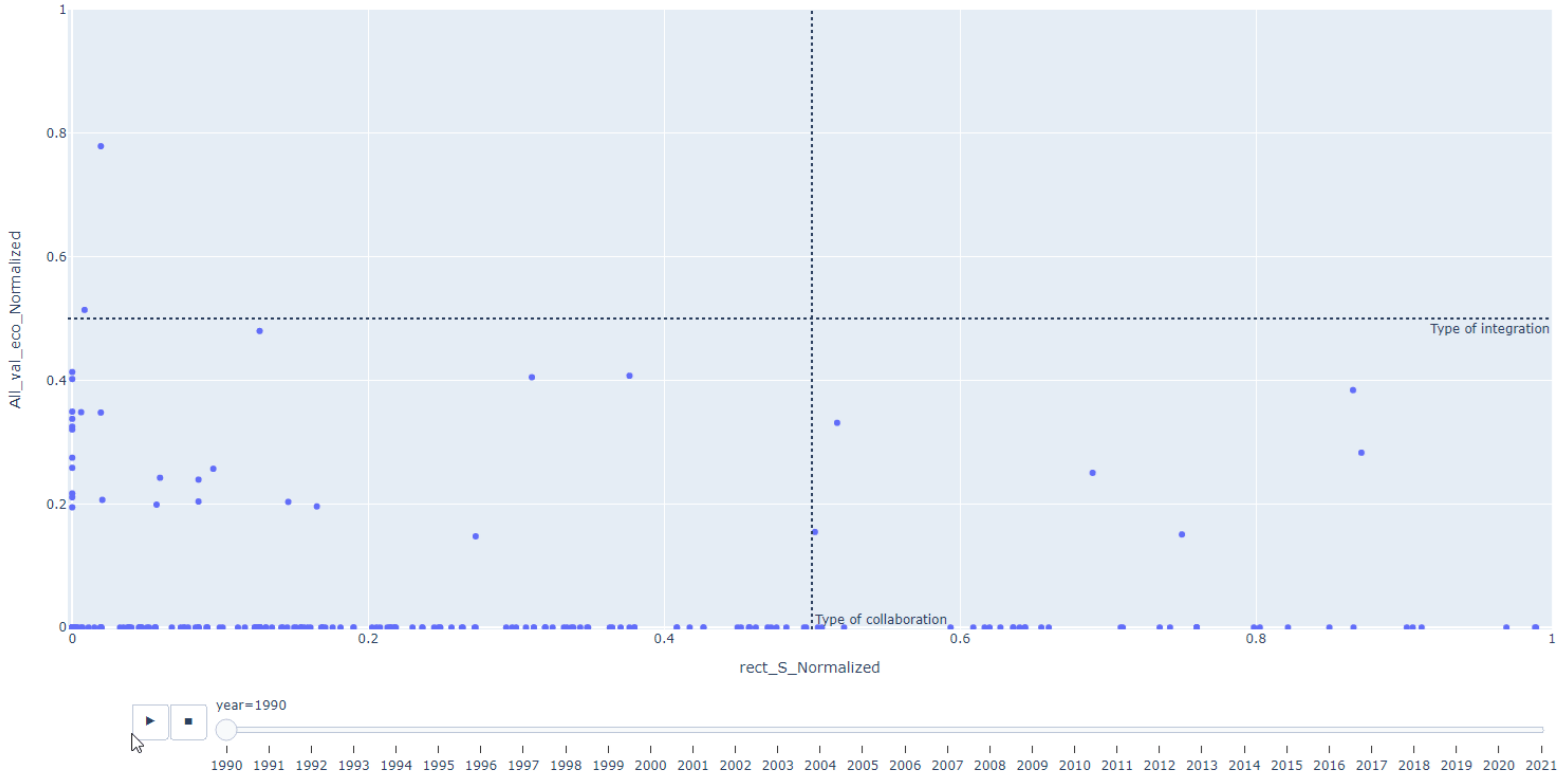



Рис. 3 – Результат 3-граммы в исследуемом тексте

Рис. 4 – статистика слова information



# Полученное распределение компаний по годам



# Оценка линейности параметров

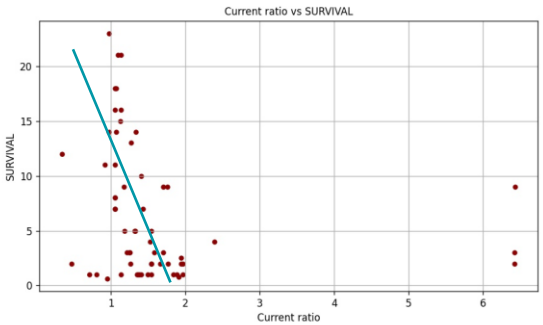



Рис. 5 - Current ratio vs Survival

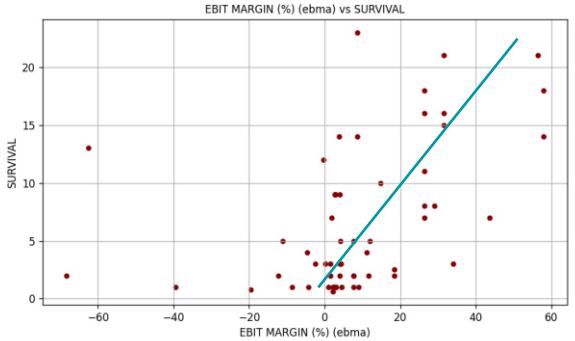



Рис. 6 - Ebit Margin (%) vs Survival

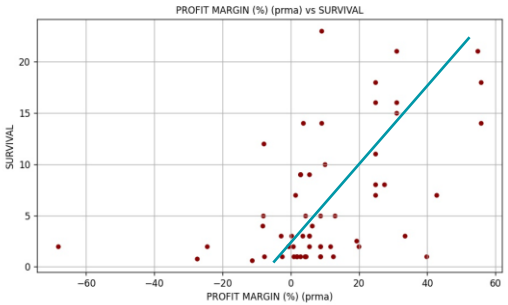



Рис. 7 - Profit Margin (%) vs Survival

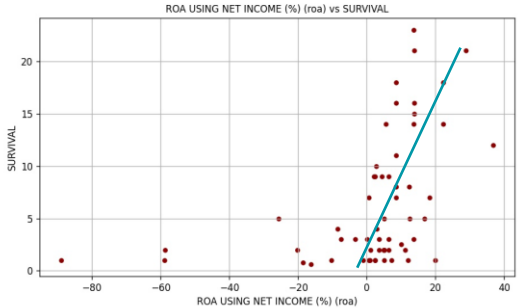
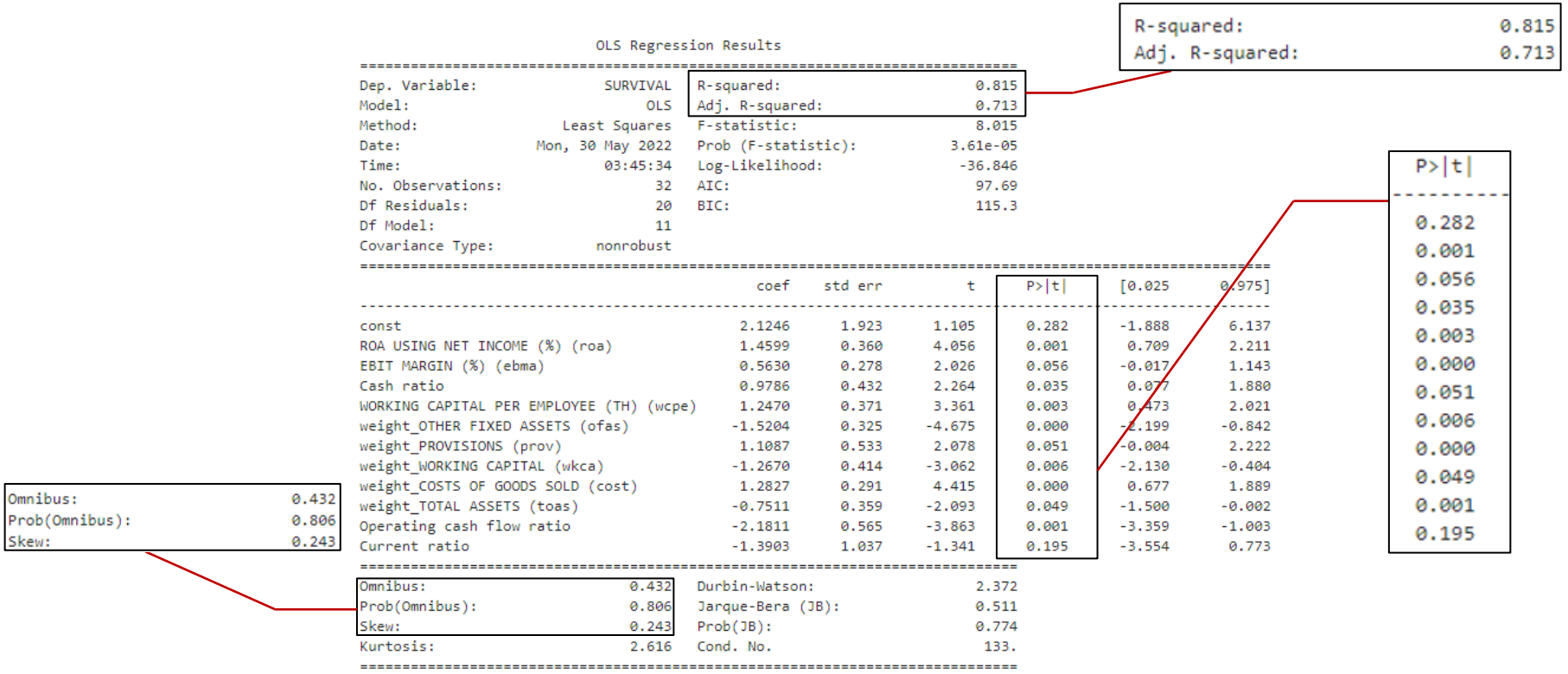



Рис. 8 - ROA Using Net Income (%) vs Survival

# Множественная линейная регрессия

Модель множественной линейной регрессии с комбинацией независимых переменных, показавших наилучший результат

$$Survival = 2.1246 + 1.4599 * ROA\ Using\ Net\ Income + 0.5630 * Ebit\ Margin + 0.9786 * Cash\ Ratio + 1.2470 * Working\ Capital\ Per\ Employee - 1.5204 * weight\ Other\ Fixed\ Assets + 1.1087 * weight\ Provisions - 1.2670 * weight\ Working\ Capital + 1.2827 * weight\ Costs\ Of\ Goods\ Sold - 0.7511 * weight\ Total\ Assets - 2.1811 * Operating\ cash\ flow\ ratio - 1.3903 * Current\ ratio + e$$




OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | SURVIVAL | R-squared: | 0.815 |
| Model: | OLS | Adj. R-squared: | 0.713 |
| Method: | Least Squares | F-statistic: | 8.015 |
| Date: | Mon, 30 May 2022 | Prob (F-statistic): | 3.61e-05 |
| Time: | 03:45:34 | Log-Likelihood: | -36.846 |
| No. Observations: | 32 | AIC: | 97.69 |
| Df Residuals: | 20 | BIC: | 115.3 |
| Df Model: | 11 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2.1246 | 1.923 | 1.105 | 0.282 | -1.888 | 6.137 |
| ROA USING NET INCOME (%) (roa) | 1.4599 | 0.360 | 4.056 | 0.001 | 0.709 | 2.211 |
| EBIT MARGIN (%) (ebma) | 0.5630 | 0.278 | 2.026 | 0.056 | -0.017 | 1.143 |
| Cash ratio | 0.9786 | 0.432 | 2.264 | 0.035 | 0.077 | 1.880 |
| WORKING CAPITAL PER EMPLOYEE (TH) (wcpe) | 1.2470 | 0.371 | 3.361 | 0.003 | 0.473 | 2.021 |
| weight_OTHER FIXED ASSETS (ofas) | -1.5204 | 0.325 | -4.675 | 0.000 | -2.199 | -0.842 |
| weight_PROVISIONS (prov) | 1.1087 | 0.533 | 2.078 | 0.051 | -0.004 | 2.222 |
| weight_WORKING CAPITAL (wkca) | -1.2670 | 0.414 | -3.062 | 0.006 | -2.130 | -0.404 |
| weight_COSTS OF GOODS SOLD (cost) | 1.2827 | 0.291 | 4.415 | 0.000 | 0.677 | 1.889 |
| weight_TOTAL ASSETS (toas) | -0.7511 | 0.359 | -2.093 | 0.049 | -1.500 | -0.002 |
| Operating cash flow ratio | -2.1811 | 0.565 | -3.863 | 0.001 | -3.359 | -1.003 |
| Current ratio | -1.3903 | 1.037 | -1.341 | 0.195 | -3.554 | 0.773 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.432 | Durbin-Watson: | 2.372 |
| Prob(Omnibus): | 0.806 | Jarque-Bera (JB): | 0.511 |
| Skew: | 0.243 | Prob(JB): | 0.774 |
| Kurtosis: | 2.616 | Cond. No. | 133. |

Рис. 9 - Наилучшая модель с набором независимых переменных, построенная в statsmodels

# Эффекты взаимодействия

Модель множественной линейной регрессии с комбинацией наилучших показателей эффектов взаимодействия и переменных

$$Survival = 8.0070 + 0.8242 * ROA\ Using\ Net\ Income + 0.8359 * Ebit\ Margin + 1.3766 * Cash\ Ratio - 1.7182 * Working\ Capital\ Per\ Employee - 1.0900 * weight\ Other\ Fixed\ Assets + 0.8907 * weight\ Provisions - 0.6683 * weight\ Working\ Capital + 0.5397 * weight\ Costs\ Of\ Goods\ Sold - 0.9301 * weight\ Total\ Assets - 2.0060 * Operating\ cash\ flow\ ratio - 4.2543 * Current\ ratio - 1.1108 * (wcpe * Cash\ ratio) + 3.8069 * (wcpe * Current\ ratio) + e$$

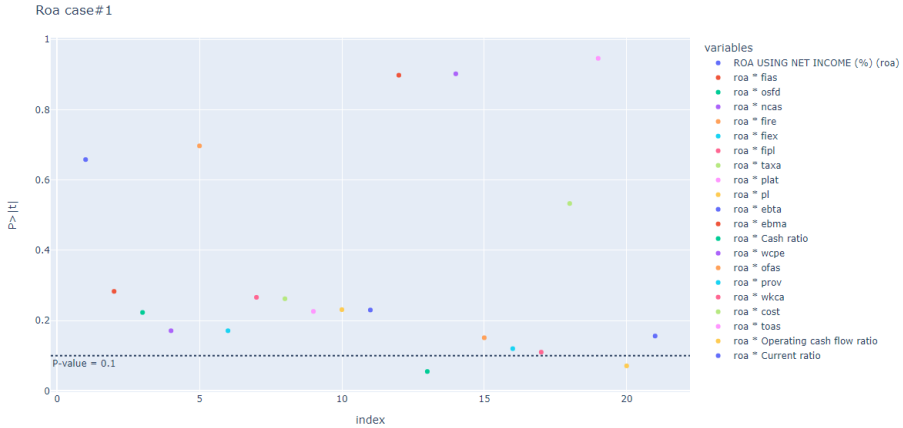
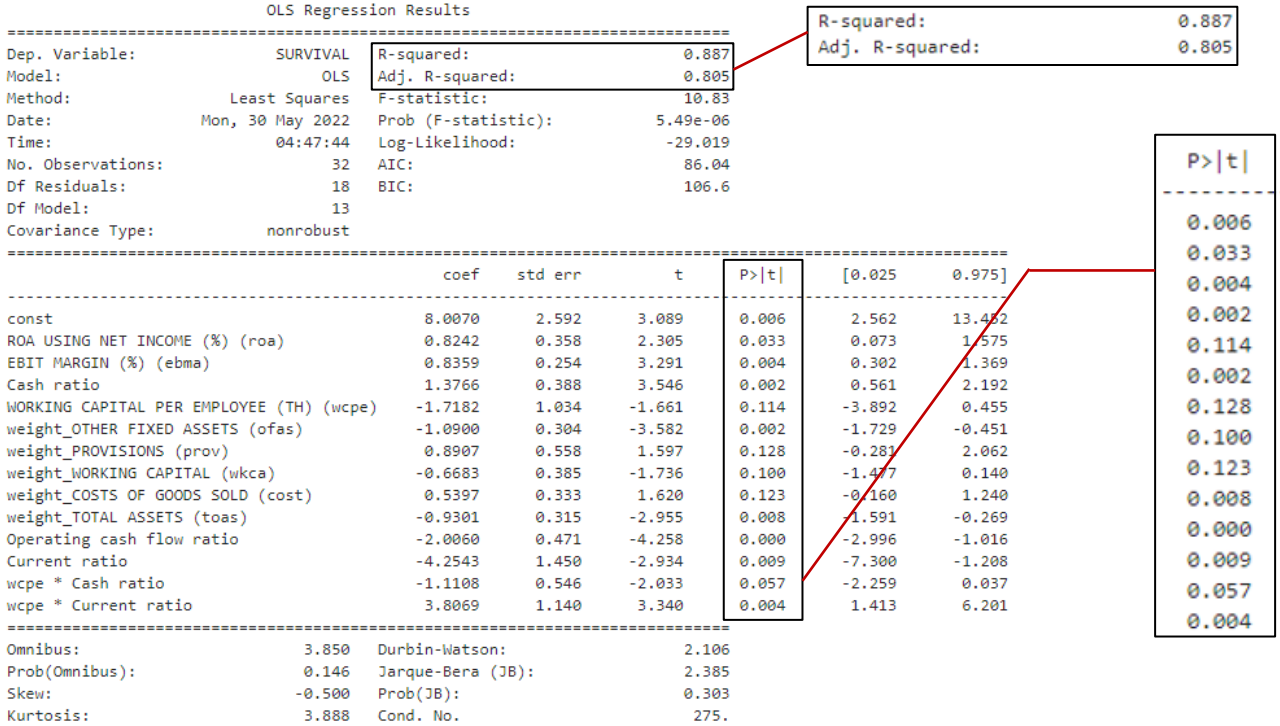



Рис. 10 – Кейс эффектов взаимодействия ROA с остальными значимыми независимыми переменными

Рис. 11 - Кейс эффектов взаимодействия WCPE с остальными значимыми независимыми переменными

# Регрессионная модель с эффектами взаимодействия



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | SURVIVAL | R-squared: | 0.887 |
| Model: | OLS | Adj. R-squared: | 0.805 |
| Method: | Least Squares | F-statistic: | 10.83 |
| Date: | Mon, 30 May 2022 | Prob (F-statistic): | 5.49e-06 |
| Time: | 04:47:44 | Log-Likelihood: | -29.019 |
| No. Observations: | 32 | AIC: | 86.04 |
| Df Residuals: | 18 | BIC: | 106.6 |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 8.0070 | 2.592 | 3.089 | 0.006 | 2.562 | 13.452 |
| ROA USING NET INCOME (%) (roa) | 0.8242 | 0.358 | 2.305 | 0.033 | 0.073 | 1.575 |
| EBIT MARGIN (%) (ebma) | 0.8359 | 0.254 | 3.291 | 0.004 | 0.302 | 1.369 |
| Cash ratio | 1.3766 | 0.388 | 3.546 | 0.002 | 0.561 | 2.192 |
| WORKING CAPITAL PER EMPLOYEE (TH) (wcpe) | -1.7182 | 1.034 | -1.661 | 0.114 | -3.892 | 0.455 |
| weight_OTHER FIXED ASSETS (ofas) | -1.0900 | 0.304 | -3.582 | 0.002 | -1.729 | -0.451 |
| weight_PROVISIONS (prov) | 0.8907 | 0.558 | 1.597 | 0.128 | -0.281 | 2.062 |
| weight_WORKING CAPITAL (wkca) | -0.6683 | 0.385 | -1.736 | 0.100 | -1.477 | 0.140 |
| weight_COSTS OF GOODS SOLD (cost) | 0.5397 | 0.333 | 1.620 | 0.123 | -0.160 | 1.240 |
| weight_TOTAL ASSETS (toas) | -0.9301 | 0.315 | -2.955 | 0.008 | -1.591 | -0.269 |
| Operating cash flow ratio | -2.0060 | 0.471 | -4.258 | 0.000 | -2.996 | -1.016 |
| Current ratio | -4.2543 | 1.450 | -2.934 | 0.009 | -7.300 | -1.208 |
| wcpe * Cash ratio | -1.1108 | 0.546 | -2.033 | 0.057 | -2.259 | 0.037 |
| wcpe * Current ratio | 3.8069 | 1.140 | 3.340 | 0.004 | 1.413 | 6.201 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.850 | Durbin-Watson: | 2.106 |
| Prob(Omnibus): | 0.146 | Jarque-Bera (JB): | 2.385 |
| Skew: | -0.500 | Prob(JB): | 0.303 |
| Kurtosis: | 3.888 | Cond. No. | 275. |

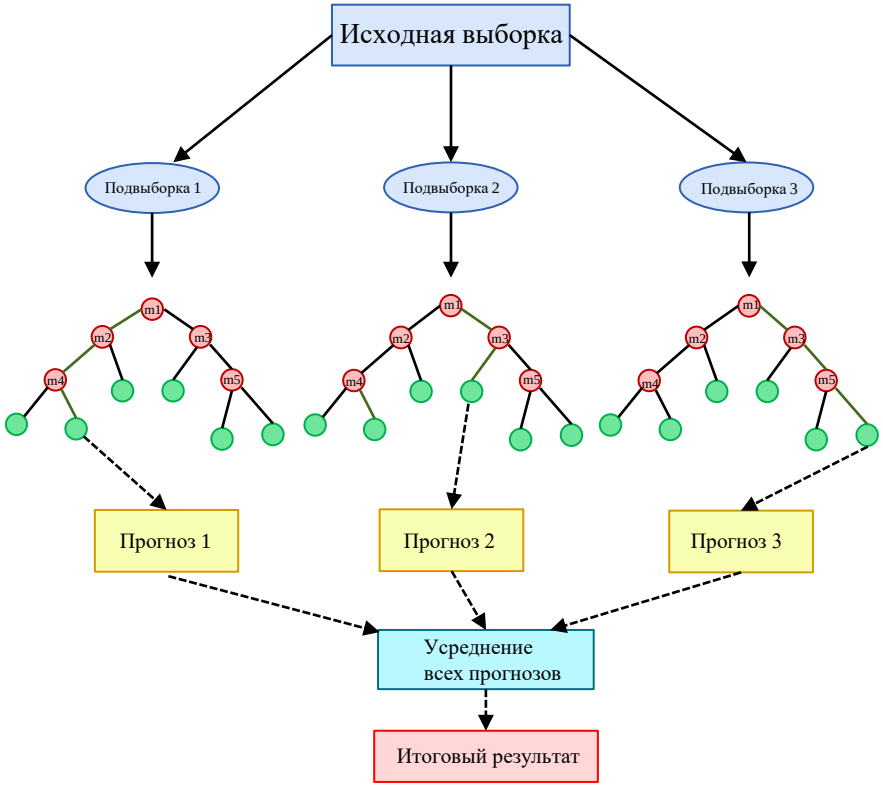


Рис. 12 - Наилучшая модель с набором независимых переменных с участием «эффектов взаимодействия», построенная в statsmodels

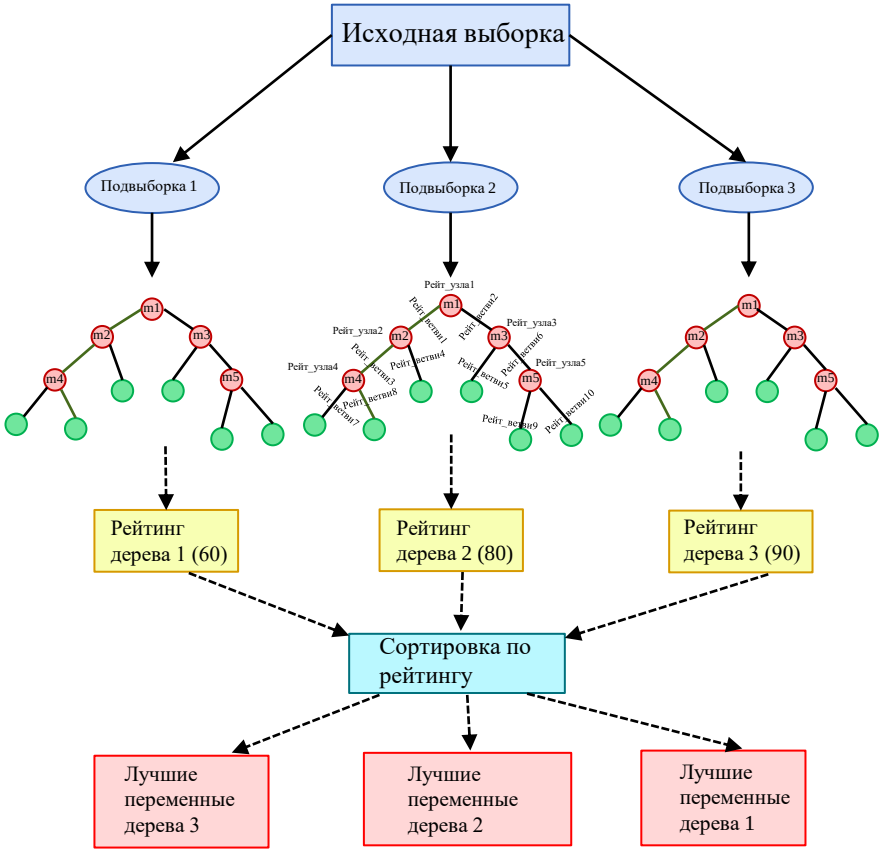
# Многоцелевой Оптимизированный Случайный Лес

- Способен обрабатывать большие наборы данных;
- Оптимизирован для дилеммы «смещения-дисперсии»;
- Демонстрация влияния независимых переменных;
- Результаты могут быть менее «хрупкими».

# Случайный лес




# Многоцелевой оптимизированный случайный лес



# Условия работы алгоритма

Каждое “дерево” в “лесу” разбивается на независимые кейсы:

- Рейтинг каждого узла (=condition)

$$Condition\ rate = 1 - \log_{10}(residual\ points)\text{, где}$$

$$residual\ points == \begin{cases} condition\ threshhold\ , & \text{если } operator = " < " \\ 10 - condition\ threshhold, & \text{если } operator = " > " \end{cases}$$

- Рейтинг каждой ветви дерева

$$Branch\ rate = \\ = \frac{correct\ predictions * \log(1 + all\ predictions)}{all\ predictions} * \sum \frac{condition\ rate}{conditions\ count}$$

- Exception = “Low”, если основная ветвь “High”. Исключение выкидывается, если рейтинг главной ветви + исключения больше, чем рейтинг просто главной ветви.

# Оценка качества

- F1 score рассчитан на основе всех кейсов тестового и тренировочного датасетов.

Каждый образец в наборе оценивается по каждому случаю. Первый совпадающий случай используется для классификации выборки.

$$F1\ score = \left(2 * \frac{(Precision\ * Recall)}{Precision + Recall}\right)$$

- Precision (точность) и coverage (охват) рассчитаны для каждого кейса отдельно на тестовой и тренировочной выборках.

$$coverage = \left(1.0 * \frac{length\ of\ results}{length\ of\ X}\right)$$

$$precision = \left(1.0 * \frac{length\ of\ matching\ results}{length\ of\ results}\right)$$

# Ручные настройки работы программы

1. Пользователь задает необходимые настройки исходя из результирующего набора данных:

   max_value = 7 - *максимальное значение, используемое в тренировочном сете данных*

   min_value = 1 - *минимальное значение, используемое в тренировочном сете данных*

   y_threshold = 3 - *порог, разделяющий верхнее и нижнее целевые значения*

   intersection_threshold = 1 - *порог для классификации случая как другого*

   min_conditions = 4 - *минимальное количество колонок для каждого кейса*

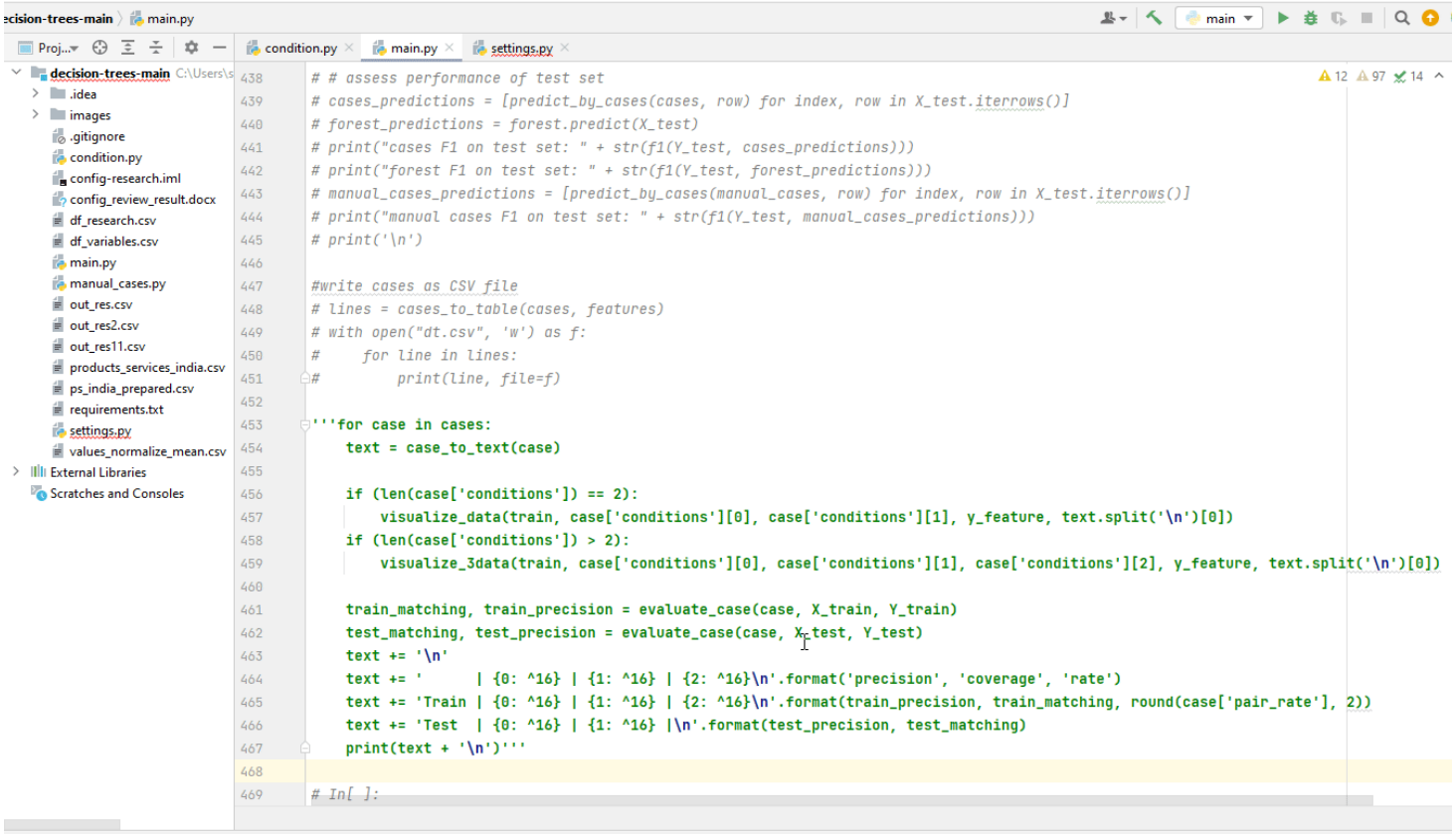
   results_count = 10 - *сколько максимально вывести деревьев в итоге*

   parsimony_rate = 0.4 - *степень экономии в рейте*

   expressiveness_rate = 0.8 - *степень выразительности в рейте*

2. Программа рандомно выбирает кейсы из датасетов и ищет наилучшие деревья решений по заданным математическим моделям;
3. Деревья решений выводятся в формате dot и визуализируются средством graph viz;
4. Также результаты выводятся в текстовом формате doc.

# Пример запуска программной реализации



# Результаты работы алгоритма в формате DOC

From tree 75: If ROA USING NET INCOME (%) (roa) is (> 1.72) and EBIT MARGIN (%) (ebma) is (< 4.81) and weight_WORKING CAPITAL (wkca) is (> 2.72) and weight_COSTS OF GOODS SOLD (cost) is (< 2.96) then value is Low

| | precision | coverage | rate |
|---|---|---|---|
| Train | 0.88 | 0.43 | 2.01 |
| Test | 0.12 | 0.57 | |

From tree 83: If weight_WORKING CAPITAL (wkca) is (> 3.88) and Cash ratio is (> 1.01) and weight_WORKING CAPITAL (wkca) is (< 6.7) and weight_PROVISIONS (prov) is (< 1.23) then value is Low

| | precision | coverage | rate |
|---|---|---|---|
| Train | 0.45 | 0.36 | 0.27 |
| Test | 0.25 | 0.29 | |

From tree 18: If WORKING CAPITAL PER EMPLOYEE (TH) (wcpe) is (< 3.92) and weight_COSTS OF GOODS SOLD (cost) is (< 1.57) and Cash ratio is (< 4.85) and WORKING CAPITAL PER EMPLOYEE (TH) (wcpe) is (> 3.14) then value is High

| | precision | coverage | rate |
|---|---|---|---|
| Train | 0.75 | 0.07 | 0.27 |
| Test | 0.0 | 0.0 | |

From tree 28: If Cash ratio is (> 1.0) and EBIT MARGIN (%) (ebma) is (< 5.31) and weight_WORKING CAPITAL (wkca) is (> 2.55) and Cash ratio is (< 3.02) then value is Low

| | precision | coverage | rate |
|---|---|---|---|
| Train | 0.74 | 0.41 | 0.18 |
| Test | 0.12 | 0.57 | |

# Результаты работы программы в формате DOT

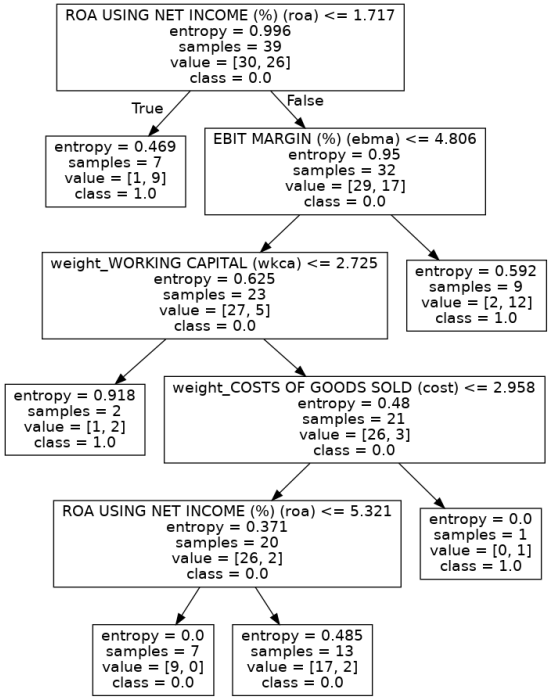



Рис. 13 - Пример отображения дерева 75 в формате DOT

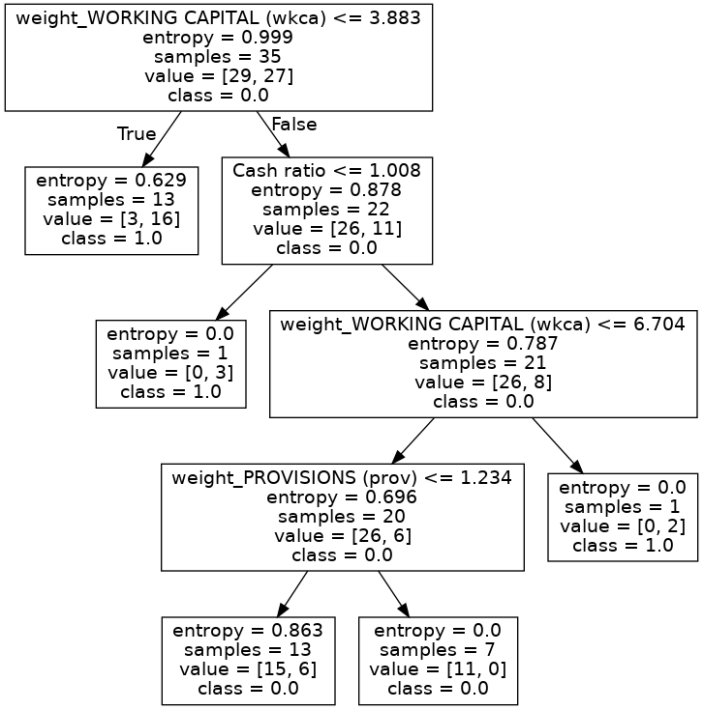



Рис. 14 - Пример отображения дерева 83 в формате DOT

# Результат исследования

Множественная линейная регрессия и Многоцелевой оптимизированный случайный лес получили очень схожие результаты и указывают на важность следующих независимых переменных, оказывающих влияние на выживаемость компаний:

- Operating cash flow ratio
- Other Fixed Assets
- Provisions
- Cash Ratio
- Working Capital Per Employee
- Costs Of Goods Sold
- ROA Using Net Income
- Total Assets
- Ebit Margin
- Working Capital
- Current ratio

# Выводы:

- Проведены очистка, исследование и анализ данных;
- Отобраны наиважнейшие независимые переменные для исследования;
- Построена множественная регрессионная модель и выбрана наилучшая комбинация независимых переменных;
- Добавлены эффекты взаимодействия;
- Реализован алгоритм МОСЛ для определения важнейших переменных, наиболее влияющих на выживаемость компаний;
- Построены деревья решений, отображающие влияние переменных на выживаемость компаний;
- Проведено сравнение результатов МОСЛ и множественной линейной регрессии;
- Полученные результаты показали правдоподобность и эффективность разработанного подхода к предсказанию выживаемости компаний.

Спасибо за внимание!

# Приложение

- **ROA Using Net Income** - термин рентабельность активов (ROA) относится к финансовому коэффициенту, который показывает, насколько прибыльна компания по отношению к ее совокупным активам.
- **Ebit Margin** - финансовый коэффициент, который измеряет прибыльность компании, рассчитанную без учета влияния процентов и налогов.
- **Cash Ratio** - мера ликвидности компании.
- **Working Capital Per Employee** - может быть связан с оборотным капиталом с точки зрения человеческих активов и затрат, связанных с работниками.
- **Other Fixed Assets** - могут включать здания, компьютерное оборудование, программное обеспечение, мебель, землю, машины и транспортные средства.
- **Provisions** - представляют собой средства, отложенные компанией для покрытия ожидаемых убытков в будущем.
- **Working Capital** - представляет собой разницу между текущими активами компании, такими как денежные средства, дебиторская задолженность/неоплаченные счета клиентов, запасы сырья и готовой продукции, и ее текущими обязательствами, такими как кредиторская задолженность и долги.
- **Costs Of Goods Sold** - это общая сумма, уплаченная бизнесом в качестве затрат, непосредственно связанных с продажей продукции.
- **Total Assets** - относятся к общей сумме активов, принадлежащих физическому или юридическому лицу.
- **Operating cash flow ratio** - это показатель того, сколько раз компания может погасить текущие долги денежными средствами, полученными в течение одного и того же периода.
- **Current ratio** - это коэффициент ликвидности, который измеряет способность компании погасить краткосрочные обязательства или обязательства со сроком погашения в течение одного года.