РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ЭМИТ
ИНСТИТУТ ЭКОНОМИКИ, МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Разработка моделей прогнозирования банкротства российских компаний с учетом их размера и отраслевой принадлежности

бакалавриат по направлению «Экономика»
программа «Экономика и Финансы»

Студент: Майорова Ксения Николаевна

Научный руководитель:
к.э.н., Полбин Андрей Владимирович

Консультант:
Фокин Никита Денисович

2021 г.

АКТУАЛЬНОСТЬ ИССЛЕДОВАНИЯ

- Кредитные учреждения нуждаются в скоринговых моделях, обладающих высокой точностью применительно к задаче прогнозирования вероятности наступления кризисной ситуации у компании, во избежание значительных потерь при предоставлении кредита компаниям
- В нестабильной рыночной среде развивающийся экономики России оценка состояния своей компании или компании-конкурента исходя из финансовых данных важно для людей, принимающих менеджерские решения для обеспечения эффективного управления компанией
- Развитие малого бизнеса влияет на экономический рост, научно-технический прогресс и расширяет число рабочих мест, а компании среднего и крупного бизнеса создают прочную основу экономического потенциала каждой страны и отличаются высокой инновационной активностью, поэтому различные государственные органы могут быть заинтересованы в прогнозировании будущей динамики финансового состояния компаний, например, с целью разработки своевременных мер поддержки бизнеса

АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

Авторы	Название работы, год	Источник	Результат
Altman E. I.	Financial ratios, discriminant analysis and the prediction of corporate bankruptcy (1968)	The journal of finance	Автор стал новатором в применении статистического инструмента MDA для оценки риска банкротства компаний на выборке 66 американских компаний в период с 1946-1965 и разработал пятифакторную модель (Z-Score Model) для публичных предприятий, чьи акции торгуются на бирже.
Ohlson J. A	Financial ratios and the probabilistic prediction of bankruptcy (1980)	Journal of accounting research	Автор впервые предложил использование логистической регрессии и разработал девятифакторную модель (O-Score), используя более 2000 наблюдений за промышленным компаниям за период 1970-1976 гг. Это положило начало массовому применению логит метода.
Демешев Б. Б., Тихонова А. С.	Прогнозирование банкротства российских компаний: межотраслевое сравнение (2014)	Экономи-ческий журнал Высшей школы экономики	Авторы моделировали критическое финансовое положение непубличных средних и малых российских компаний в 2011–2012 гг. с межотраслевым сравнением и использовали семь методов: ЛДА, КДА, СДА, классификационные деревья, алгоритм случайного леса, логит- и пробит-модели. Вне зависимости от отрасли наилучшим методом оказался алгоритм случайного леса. Предельные эффекты логит-модели по отраслям показали, что отрасли довольно сильно отличаются друг от друга.

АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

Авторы	Название работы, год	Источник	Результат
Казаков А. В., Колышкин А. В.	Разработка моделей прогнозирования банкротства в современных российских условиях (2018)	Вестник Санкт-Петербургского университета. Экономика.	На основе наблюдений за компаниями различных отраслей в 2014–2015 гг. авторы оценили 35 зарубежных и отечественных моделей, среди которых как классические, так и популярные последних лет. Ни одна из моделей ни в одной отрасли его не превзошла точность 70%. Далее авторы разработали собственную модель для каждой отрасли и на тестовой выборке все модели продемонстрировали точность в среднем 70%, то есть они оказались более устойчивы, и межотраслевая классификация позволила добиться более высокой точности при прогнозе.
Fedorova E., Gilenko E., Dovzhenko S.	Bankruptcy prediction for Russian companies: Application of combined classifiers (2013)	Expert systems with applications	Авторы применили более современные подходы, используя наблюдения по крупным и средним российским компаниям в период 2007–2011 гг. Среди классических моделей модель Фулмера показала самую высокую общую точность (82%). Также авторы применили два типа нейронных сетей: многослойный перцептрон (MLP) и сеть радиально-базисных функций (RBFN). Для них был проведен процесс предварительного отбора переменных. В итоге MLP, построенный с помощью отобранных деревром решений переменных, продемонстрировал самые высокие результаты. Для объединения и комбинирования результатов была применена методология AdaBoost. Итоговая общая точность составила 88,8%, что доказывает необходимость применять современные методики для разработки более эффективного классификатора.

ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

ЦЕЛЬ ИССЛЕДОВАНИЯ:

Разработка высокоточного классификатора с целью прогнозирования дефолта российских компаний с учётом их размера и отраслевой принадлежности с помощью различных методов эконометрики и машинного обучения, используя финансовые показатели из бухгалтерской отчетности компании, а также некоторые её нефинансовые характеристики.

ЗАДАЧИ ИССЛЕДОВАНИЯ:

- Осуществить сбор и первичную обработку данных из базы данных РУСЛАНА
- Имплементировать программный код на языке Python
- Проанализировать полученные результаты и сделать основные выводы о практической значимости исследования

Сбор данных

В выборке содержатся компании, которые обанкротились в 2018 и в 2019 годах. Финансовая отчетность для объясняющих переменных берется за год до банкротства.

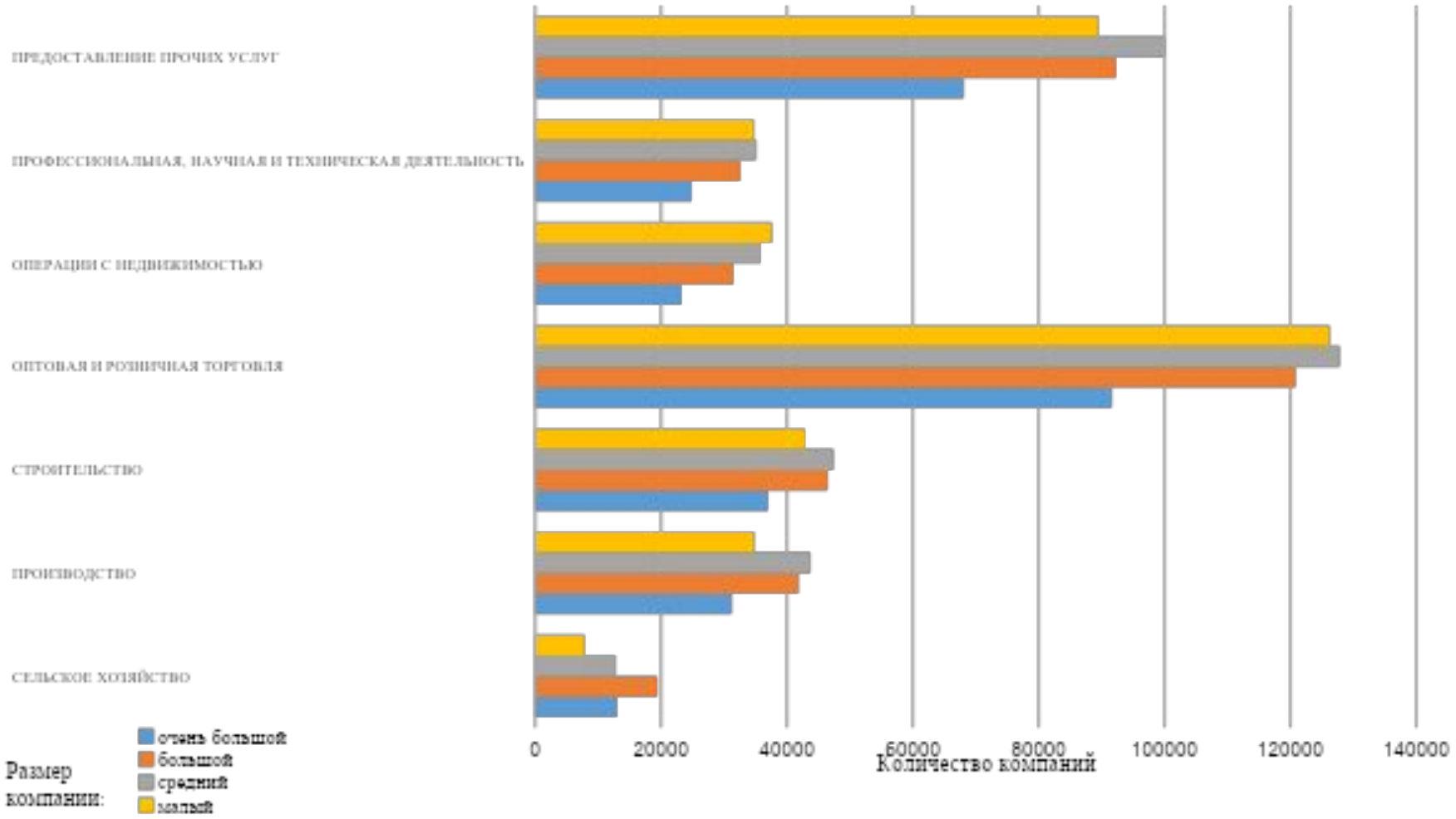
	Действующие компании	Банкроты
Изначально всего	~ 4 млн. 200 тыс.	10 759
Те, у которых нет пропущенных значений в переменных	~ 1 млн. 550 тыс.	9 982
Обучение / тест (пропорция 0.8 / 0.2)*	7 985 / 287 988	7 985 / 1 997

*Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. Knowledge-Based Systems, 41, 16-25.

Объясняющих переменных получилось более 50:

1. Финансовые переменные взяты из классических зарубежных и отечественных моделей и из обзора литературы.
2. Нефинансовые переменные: возраст компании, число работников, дамми на размер, дамми на отрасль, дамми на импортную и экспортную деятельность.

Распределение компаний в выборке в зависимости от размера и отрасли



Отобранные для разработки классификатора модели

Модель	Предикторы	Финансовые показатели
Модель Альтмана-Собато	1. EBIT/A 2. SL/EQ 3. RI/A 4. CASH/A 5. EBIT/IP	A - активы; CA - оборотные активы; CASH - денежные средства; EQ - собственный капитал; RI - нераспределенная прибыль; L - общие обязательства; LL - долгосрочные обязательства; SL - краткосрочные обязательства; S - выручка от продажи; IP - проценты к уплате; NP - чистая прибыль; WC - собственные оборотные средства (превышение оборотных активов над краткосрочными обязательствами); EBIT - прибыль до вычета процентов и налогов.
Модель Фулмера	1. RI/A 2. S/A 3. EBIT/EQ 4. CASH/L 5. LL/A 6. L/A 7. log_A 8. WC/L 9. log_EBIT/IP	
Модель Лиса	1. RI/A 2. EBIT/A 3. EQ/L 4. WC/A	
Модель Змиевского	1. L/A 2. NP/A 3. CA/SL	
Модель Алексеевой	1. S/A 2. NP/A 3. L/A 4. LL/A 5. log_S	

Помимо этого, были отобраны несколько методов машинного обучения и эконометрики, а именно, логистическая регрессия с L1-регуляризацией (Logit-Lasso), дерево решений (Decision Tree), случайный лес (Random Forest), градиентный бустинг (Gradient Boosting) и ставший популярным недавно разработанный компанией Яндекс усовершенствованный метод градиентного бустинга CatBoost.

Описание эксперимента и сравнение моделей по качеству прогноза

Имелось 10 видов моделей (Альтмана-Собато, Фулмера, Лиса, Змиевского, Алексеевой, Logit-Lasso, Decision Tree, Random Forest, Gradient Boosting, CatBoost) и 4 вида выборок, когда для обучения этих моделей используется:

1) вся выборка целиком (10 моделей);
2) выборки с компаниями одного размера (4*10 = 40 моделей);
3) выборки с компаниями одной отрасли (7*10 = 70 моделей);
4) выборки с компаниями и одного размера, и одной отрасли (4*7*10 = 280 моделей).

Качество моделей при прогнозе и его изменение в зависимости от типа обучения (в 1-й строке представлен ROC AUC, в остальных строках – его процентное изменение по отношению к общей модели, полужирным выделено наилучшее улучшение, красным – случаи ухудшения качества прогноза)

Вид обуча-ющей выборки	Мо-дель	Альтман-Собато	Фулмер	Лис	Змиевский	Алексеева	Logit-Lasso	Decision Tree	Random Forest	CatBoost	Gradient Boosting
Одна общая модель		0.702	0.722	0.685	0.711	0.587	0.756	0.791	0.820	0.833	0.828
Отдельная модель для каждого размера		**4.3%**	3.8%	**1.8%**	-2.8%	15.2%	2.6%	**1.7%**	**0.8%**	-0.1%	**0.2%**
Отдельная модель для каждой отрасли		-1.6%	3.7%	-1.5%	-5.7%	7.0%	**2.8%**	-0.3%	-0.4%	-0.7%	-0.5%
Отдельная модель и для размера, и для отрасли		1.1%	**5.4%**	0.1%	-3.8%	**16.9%**	2.3%	-2.0%	-0.4%	-1.5%	-0.5%

Предельные эффекты моделей Logit-Lasso для компаний разных типов размеров

Предикторы	Значение	Очень крупные компании	Крупные компании	Средние компании	Мелкие компании
	возраст компании	-0.021 **(0.000)**	0.021 **(0.000)**	-0.036 **(0.000)**	-0.008 **(0.004)**
	среднесписочная численность работников		0.0006 **(0.003)**		0.007 **(0.000)**
	чистая прибыль	-0.00004 **(0.000)**	-0.0004 **(0.002)**		
	нераспределенная прибыль/активы	-0.330 **(0.000)**	-0.033 **(0.005)**		
	денежные средства/общие обязательства	-0.009 **(0.000)**	-0.4274 **(0.000)**		-0.019 **(0.011)**
	общие обязательства/активы	-0.2124 **(0.000)**		-0.0122 **(0.000)**	-0.300 **(0.000)**
	натуральный логарифм активов	0.062 **(0.000)**		0.1135 **(0.000)**	0.084 **(0.000)**
	собственные оборотные средства/общие обязательства	0.003 **(0.002)**	-0.001 (0.063)		0.00005 (0.912)
	логарифм отношения прибыли до вычета процентов и налогов к процентам к уплате	-0.063 **(0.000)**			
	собственный капитал/общие обязательства	-0.003 **(0.001)**			
	собственные оборотные средства/активы	-0.082 **(0.000)**	0.029 **(0.016)**	-0.065 **(0.000)**	-0.301 **(0.000)**
	прибыль до вычета процентов и налогов		0.0003 **(0.049)**		
	натуральный логарифм выручки от продажи		-0.020 **(0.000)**	-0.057 **(0.000)**	-0.027 **(0.000)**
	денежные средства/активы			-0.865 **(0.000)**	-0.217 **(0.000)**
	чистая прибыль/активы			-0.041 **(0.000)**	
	отрасль производства			-0.090 **(0.001)**	
	отрасль строительства			-0.160 **(0.000)**	-0.015 (0.544)
	долгосрочные обязательства/активы				0.299

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ЭМИТ
ИНСТИТУТ ЭКОНОМИКИ, МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Предельные эффекты моделей Logit-Lasso для компаний разных типов размеров

Предикторы	Сельское хозяйство	Производство	Строительство	Торговля	Недвижимость	Научная деятельность	Прочие услуги
возраст компании	-0.022 **(0.000)**	0.020 **(0.000)**	0.012 **(0.000)**	-0.033 **(0.000)**	-0.047 **(0.000)**	-0.043 **(0.000)**	-0.017 **(0.000)**
среднесписочная численность работников				0.0007 **(0.000)**	0.0002 (0.581)		
чистая прибыль				-0.0001 **(0.000)**	-0.0005 **(0.006)**		-0.00002 (0.425)
собственный капитал					-0.00002 **(0.000)**		
нераспределенная прибыль/активы	-0.079 **(0.005)**		-0.017 (0.094)				
денежные средства/общие обязательства		-0.445 **(0.000)**	-0.471 **(0.000)**	-0.059 **(0.001)**	-0.019 **(0.015)**	0.012 **(0.002)**	-0.8247 **(0.000)**
денежные средства		-0.00001 (0.392)		-0.00008 **(0.016)**			
выручка от продажи/активы	-0.0003 (0.828)				-0.000003 (0.994)		
натуральный логарифм активов	0.093 **(0.000)**			0.127 **(0.000)**	0.064 **(0.000)**	0.076 **(0.000)**	0.0809 **(0.000)**
собственные оборотные средства/общие обязательства	-0.072 **(0.000)**	-0.001 **(0.009)**	-0.006 (0.068)	0.004 (0.238)		0.005 (0.588)	0.0004 (0.734)
прибыль до вычета процентов и налогов/ собственный капитал	-0.016 (0.057)			-0.001 **(0.046)**			-0.0014 (0.135)
собственный капитал/общие обязательства		-0.036 **(0.006)**	-0.0007 (0.604)	-0.008 **(0.006)**		-0.017 **(0.035)**	-0.0017 (0.097)
собственные оборотные средства/общие обязательства	0.0790 **(0.005)**	-0.006 **(0.033)**	-0.028 **(0.019)**	-0.005 **(0.004)**		-0.0003 **(0.038)**	-0.0002 (0.348)
оборотные активы/ краткосрочные обязательства	-0.011 **(0.016)**	-0.0003 (0.678)		-0.001 **(0.018)**			
прибыль до вычета процентов и налогов					0.0004 **(0.019)**		0.00001 (0.695)
натуральный логарифм выручки от продажи	-0.062 **(0.000)**	-0.007 **(0.035)**	-0.0036 (0.156)	-0.082 **(0.000)**			-0.041 **(0.000)**
чистая прибыль/активы							-0.001 (0.151)
долгосрочные обязательства/активы	-0.058 (0.057)	-0.007 (0.360)					

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ЭМИТ
ИНСТИТУТ ЭКОНОМИКИ, МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

РЕЗУЛЬТАТ ИССЛЕДОВАНИЯ

РЕЗУЛЬТАТ:

Был разработан высокоточный классификатор с целью прогнозирования дефолта российских компаний с учётом их размера и отраслевой принадлежности, а также проведен эксперимент для выявления, помогает ли построение различных моделей для групп компаний по типу размера, отрасли, отрасли и размера одновременно получить более качественный результат.

ВЫВОДЫ:

- При построении общих моделей методы машинного обучения, особенно CatBoost, показали качество более высокое, чем классические линейные модели;
- Согласно проведенному эксперименту, для 8 из 10 моделей качество прогноза улучшилось, если для каждого отдельного типа предприятия по размеру (очень крупное, крупное, среднее, малое) строить отдельную модель, а не общую;
- Разбивка обучающей выборки на предприятия по категориям показывает улучшение прогноза лишь в 3 и 5 случаев, что ставит под сомнение выгоду и пользу применения такого подхода, встречавшегося в более ранних отечественных работах;
- В среднем, прирост качества от построения отдельных моделей выше для более слабых классических моделей, чем для более сильных моделей машинного обучения;
- Прогноз, построенный методом Logistic Lasso возможно улучшить любым из трех предложенных подходов обучения;
- При построении Logistic Lasso в зависимости от размера компании и отрасли в целом множество переменных часто выбирается схожим, однако предельные эффекты разняться и по абсолютному значению, и по знаку влияния.

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ЭМИТ
ИНСТИТУТ
ЭКОНОМИКИ, МАТЕМАТИКИ
И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Спасибо за внимание!