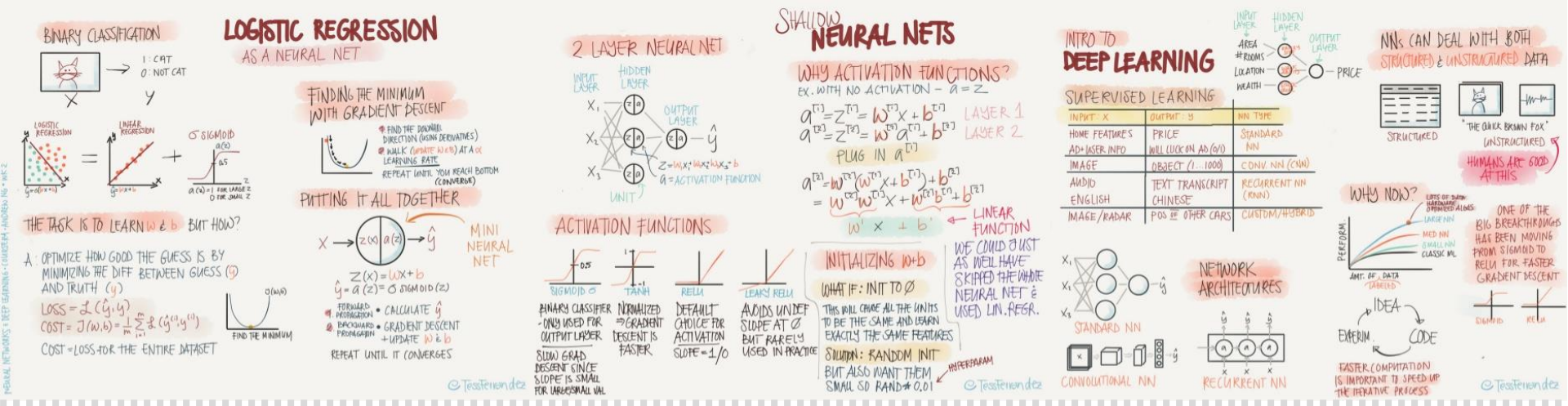


Предсказание дохода заемщика

Практический семинар

Сергей Афанасьев, Светлана Малько
КБ «Ренессанс Кредит»

13 марта 2019 г.
МШЭ МГУ, Москва

Коротко о нас и наших планах

Афанасьев Сергей

Начальник управления статистического анализа

Начальник управления расследования мошенничества

Малько Светлана

Начальник отдела разработки и анализа эффективности скоринговых систем

10:40 - 11:20

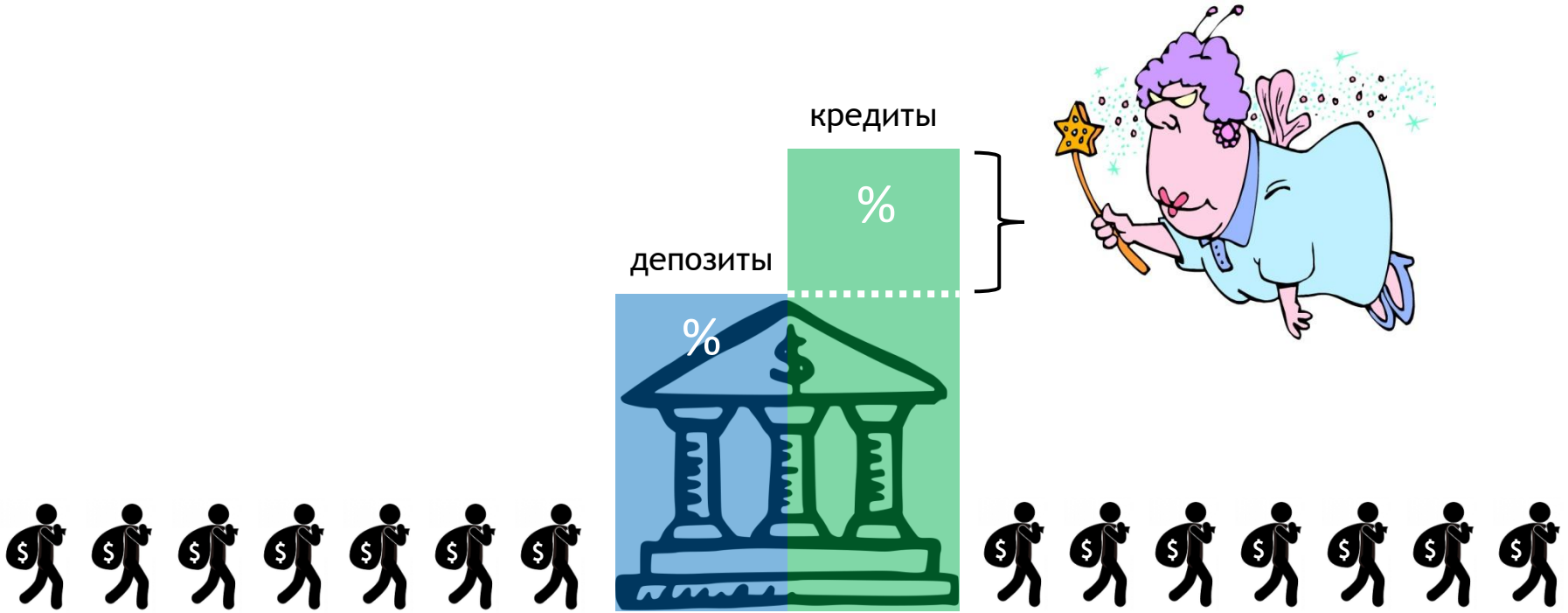
- Как устроен банковский бизнес?
- Задача предсказания дохода клиента: постановка и решение

11:30 - 12:10

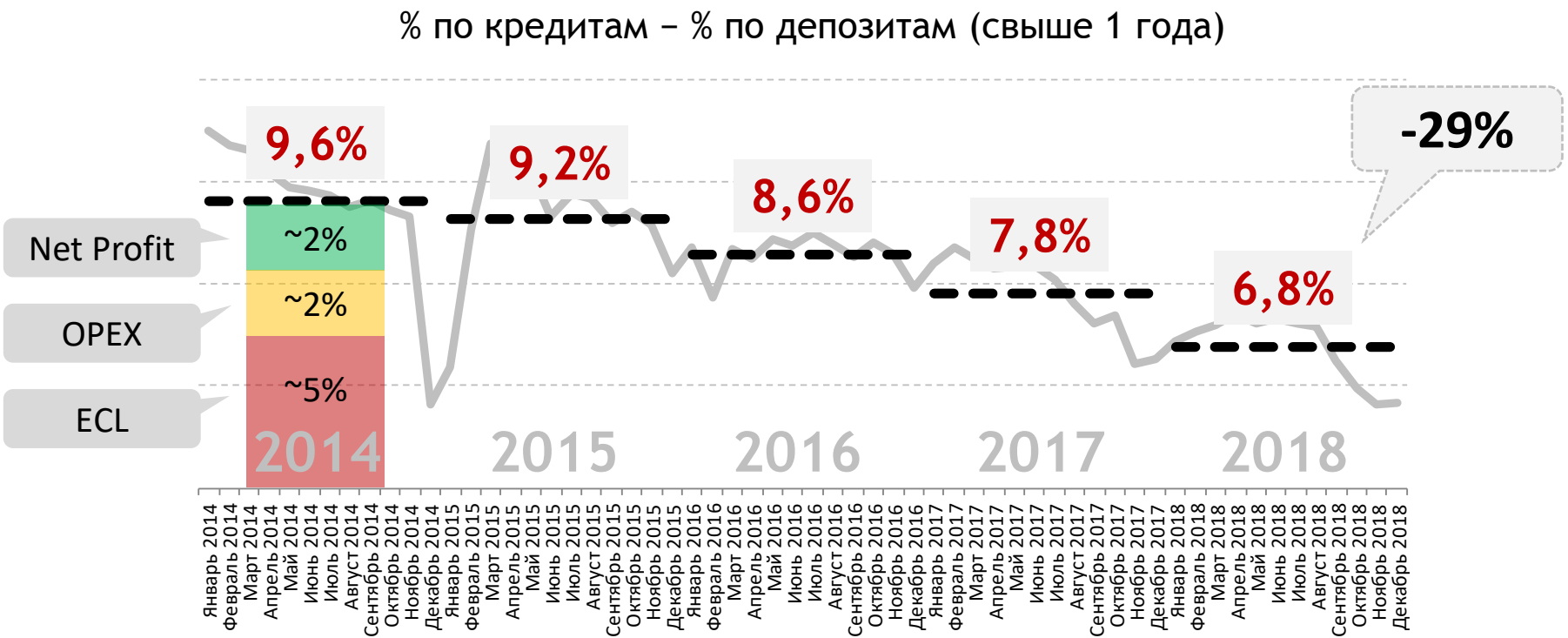
- Решение задачи в R с помощью многомерной регрессии

Вопросы и ответы

Как устроен банковский бизнес



Процентная маржа снижается



Источник: www.cbr.ru

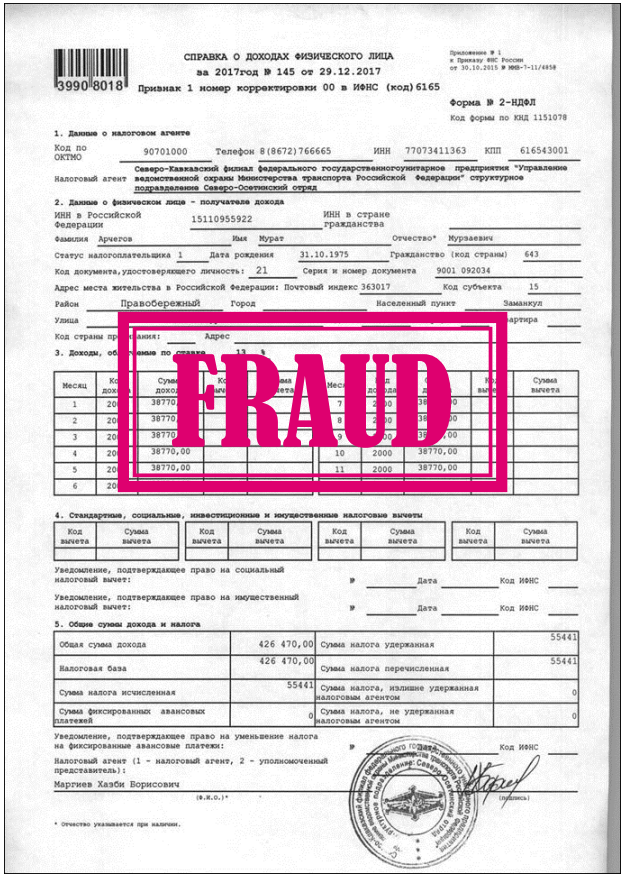
Data Science в банке

	Risk, Antifraud	Collection	CRM	CC, TM, CS	IT, IT Security	Other
Classical Machine Learning	**Credit Scoring** - Application scoring - Behavioral scoring **Anti-fraud Models** - Внутренний фрод - Внешний фрод - Транзакционный фрод	**Collection Scoring** - Модели Recovery Rate	**Recommender System** - Модели отклика на коммуникацию **Targeted Advertising** - Таргетирование рекламы на сайте банка **Churn Prediction** - Модели оттока		**Biometrics** - Keystroke Dynamics & Mouse Movements - Device Print для ИБ/МБ **Error Detection** - Выявление сбоев в системах Банка	**AML** - AML правила и модели **Time Normalization** - Оптимальная загрузка массовых подразделений **Staff Recruitment** - Отбор резюме по ключевым словам
Natural Language Processing	**Voice Recognition** - Аутентификация клиентов по голосу **Fraud Text Mining** - Анализ корпоративной переписки	**Text-To-Speech** - Голосовой коллектор	**Recommender System** - Анализ транзакции с помощью RNN для построения моделей отклика	**Speech recognition** **CS Text Mining** - Извлечение инфо - Анализ тональности - Чат-боты (сайт, ИБ/МБ) **Text-To-Speech** - Синтез речи	**Security Text Mining** - Проверка email-писем на утечку информации **Topic Model** - Распределение заявок в HD по тематикам	**Post Mail Classifier** - Распределение почты по тематикам **HR Text Mining** - Анализ переписки и выявление проблем в работе сотрудников
Computer Vision	**Photo Biometrics** - Вход в ИБ/МБ - Вход в системы Банка **Object Detection** - Скоринг по фото - Проверка фото ТТ **Spoofing** - Выявление подделок			**OCR (Text recognition)** - Перевод скан.копий документов в текст.		**Photo Biometrics** - Аутентификация клиентов в ДО/ККО - Пропускные системы для сотрудников

Некоторые требования регулятора

1. **Формирование резервов под кредитные потери (дефолты)**
2. **Ограничения на ежемесячные платежи клиентов (PTI)**
3. **Подтверждение дохода клиента (справки 2-НДФЛ и т.п.)**

Справка о доходе 2-НДФЛ



СПРАВКА О ДОХОДАХ ФИЗИЧЕСКОГО ЛИЦА
за 2017год № 145 от 29.12.2017
Признак 1 номер корректировки 00 в ИФНС (код) 6165

Форма № 2-НДФЛ
Код формы по КНД 1151078

1. Данные о налоговом агенте
Код по ОКТМО 90701000 Телефон 8(8672)766665 ИНН 77073411363 КПП 616543001
Налоговый агент Северо-Кавказский филиал федерального государственного унитарного предприятия "Управление ведомственной охраны Министерства транспорта Российской Федерации" структурное подразделение Северо-Осетинский отряд

2. Данные о физическом лице – получателе дохода
ИНН в Российской Федерации 15110955922 ИНН в стране гражданства
Фамилия Арчегов Имя Мурат Отчество* Мурзаевич
Статус налогоплательщика 1 Дата рождения 31.10.1975 Гражданство (код страны) 643
Код документа,удостоверяющего личность: 21 Серия и номер документа 9001 092034
Адрес места жительства в Российской Федерации: Почтовый индекс 363017 Код субъекта 15
Район Правобережный Город Населенный пункт Заманкул
Улица Квартира
Код страны проживания: Адрес

3. Доходы, облагаемые по ставке 13 %

Месяц	Код дохода	Сумма дохода	Код вычета	Сумма вычета	Месяц	Код дохода	Сумма дохода	Код вычета	Сумма вычета
1	2000	38770,00			7	2000	38770,00		
2	2000	38770,00			8	2000	38770,00		
3	2000	38770,00			9	2000	38770,00		
4	2000	38770,00			10	2000	38770,00		
5	2000	38770,00			11	2000	38770,00		
6	2000				12				

FRAUD

4. Стандартные, социальные, инвестиционные и имущественные налоговые вычеты

Код вычета	Сумма вычета	Код вычета	Сумма вычета	Код вычета	Сумма вычета	Код вычета	Сумма вычета

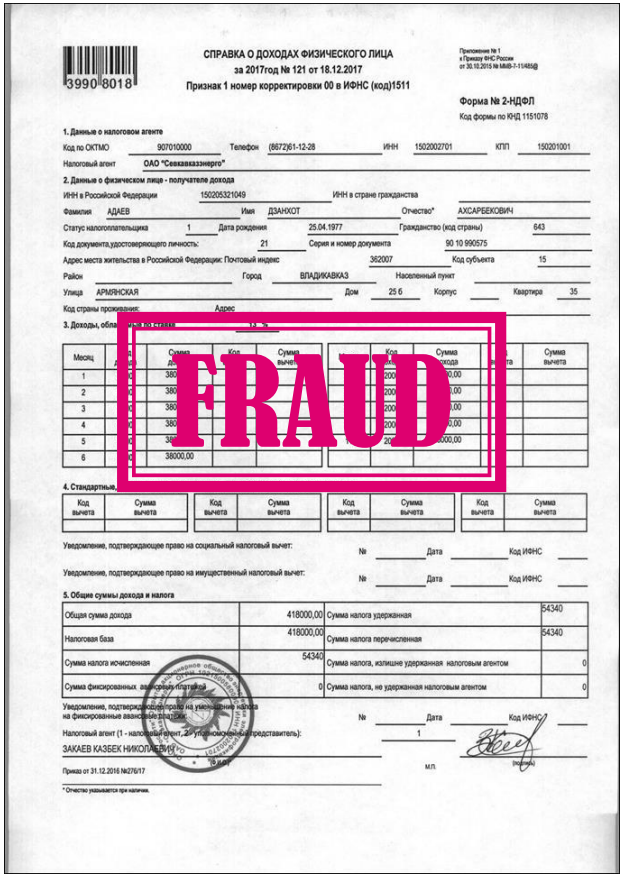
Уведомление, подтверждающее право на социальный налоговый вычет: № Дата Код ИФНС
Уведомление, подтверждающее право на имущественный налоговый вычет: № Дата Код ИФНС

5. Общие суммы дохода и налога

Общая сумма дохода	426 470,00	Сумма налога удержанная	55441
Налоговая база	426 470,00	Сумма налога перечисленная	55441
Сумма налога исчисленная	55441	Сумма налога, излишне удержанная налоговым агентом	0
Сумма фиксированных авансовых платежей	0	Сумма налога, не удержанная налоговым агентом	0

Уведомление, подтверждающее право на уменьшение налога на фиксированные авансовые платежи: № Дата Код ИФНС
Налоговый агент (1 - налоговый агент, 2 - уполномоченный представитель):
Мартиев Хазби Борисович
(Ф.И.О.)* (подпись)

* Отчество указывается при наличии.



СПРАВКА О ДОХОДАХ ФИЗИЧЕСКОГО ЛИЦА
за 2017год № 121 от 18.12.2017
Признак 1 номер корректировки 00 в ИФНС (код)1511

Форма № 2-НДФЛ
Код формы по КНД 1151078

1. Данные о налоговом агенте
Код по ОКТМО 907010000 Телефон (8672)61-12-26 ИНН 1502002701 КПП 150201001
Налоговый агент ОАО "Севкавказэнерго"

2. Данные о физическом лице - получателе дохода
ИНН в Российской Федерации 150205321049 ИНН в стране гражданства
Фамилия АДАЕВ Имя ДЗАНХОТ Отчество* АХСАРБЕКОВИЧ
Статус налогоплательщика 1 Дата рождения 25.04.1977 Гражданство (код страны) 643
Код документа,удостоверяющего личность: 21 Серия и номер документа 90 10 990575
Адрес места жительства в Российской Федерации: Почтовый индекс 362007 Код субъекта 15
Район Город ВЛАДИКАВКАЗ Населенный пункт
Улица АРМЯНСКАЯ Дом 25 б Корпус Квартира 35
Код страны проживания: Адрес

3. Доходы, облагаемые по ставке 13 %

Месяц	Код дохода	Сумма дохода	Код вычета	Сумма вычета	Месяц	Код дохода	Сумма дохода	Код вычета	Сумма вычета
1	2000	38000,00			7	2000	38000,00		
2	2000	38000,00			8	2000	38000,00		
3	2000	38000,00			9	2000	38000,00		
4	2000	38000,00			10	2000	38000,00		
5	2000	38000,00			11	2000	38000,00		
6	2000	38000,00			12				

FRAUD

4. Стандартные, социальные, инвестиционные и имущественные налоговые вычеты

Код вычета	Сумма вычета	Код вычета	Сумма вычета	Код вычета	Сумма вычета	Код вычета	Сумма вычета

Уведомление, подтверждающее право на социальный налоговый вычет: № Дата Код ИФНС
Уведомление, подтверждающее право на имущественный налоговый вычет: № Дата Код ИФНС

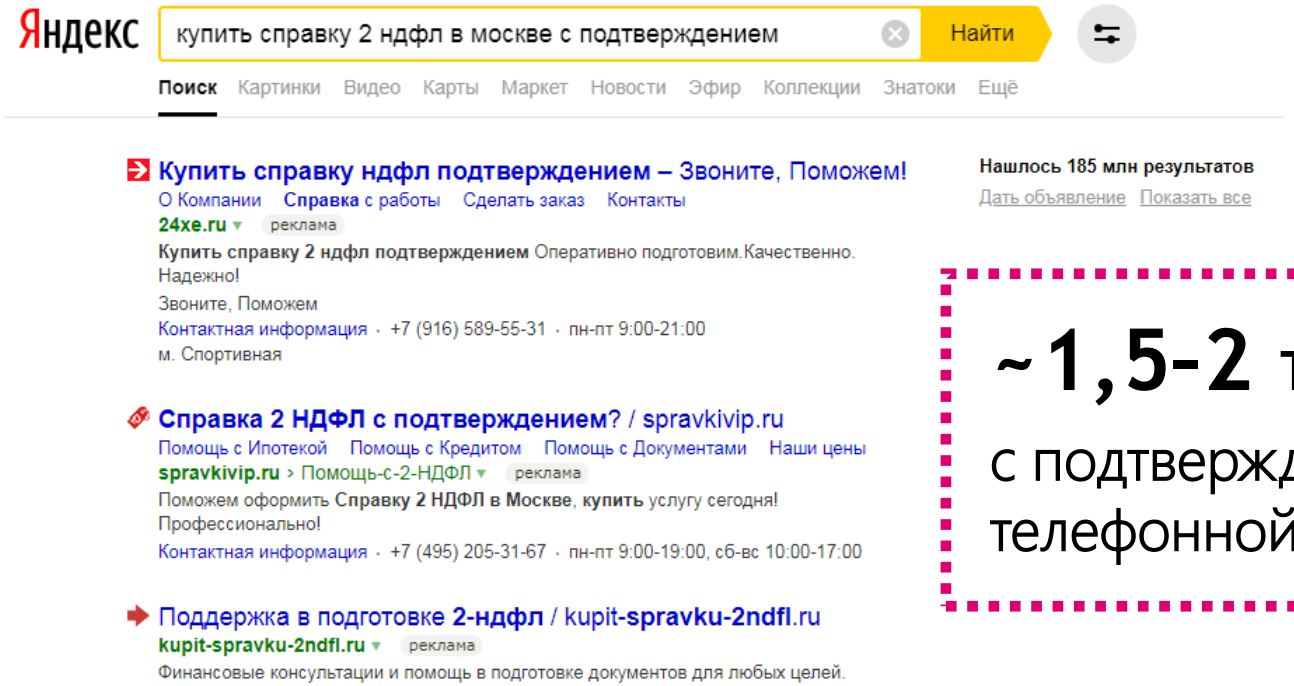
5. Общие суммы дохода и налога

Общая сумма дохода	418000,00	Сумма налога удержанная	54340
Налоговая база	418000,00	Сумма налога перечисленная	54340
Сумма налога исчисленная	54340	Сумма налога, излишне удержанная налоговым агентом	0
Сумма фиксированных авансовых платежей	0	Сумма налога, не удержанная налоговым агентом	0

Уведомление, подтверждающее право на уменьшение налога на фиксированные авансовые платежи: № Дата Код ИФНС
Налоговый агент (1 - налоговый агент, 2 - уполномоченный представитель): 1
ЗАКАЕВ КАЗБЕК НИКОЛАЕВИЧ
(Ф.И.О.)* м.п. (подпись)
Приказ от 31.12.2016 №276/17

* Отчество указывается при наличии.

2-НДФЛ увеличивают риски



~1,5-2 тыс. руб.

с подтверждением при телефонной проверке!

Многомерная регрессия

Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ — числовые признаки;

Модель многомерной линейной регрессии:

$$f(x,\alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \qquad \alpha \in R^n$$

Матричные обозначения:

$$\underset{l\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & & f_n(x_l) \end{pmatrix}, \quad \underset{l\times 1}{y} = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix}, \quad \underset{n\times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_l \end{pmatrix}$$

Функционал квадрата ошибки:

$$Q(\alpha, X^l) = \sum_{i=1}^{l} \left(f(x_i,\alpha) - y_i\right)^2 = \|F\alpha - y\|^2 \to \min_{\alpha}$$

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^T(F\alpha - y) = 0,$$

откуда следует нормальная система задачи МНК:

$$F^T F\alpha = F^T y,$$

где $F^T F$ — $n \times n$-матрица

Решение системы: $\alpha^* = \left(F^T F\right)^{-1} F^T y = F^+ y$.

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,

где $P_F = FF^+ = F\left(F^T F\right)^{-1} F^T$ — проекционная матрица.

Сингулярное разложение

Произвольная $l \times n$-матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^T$$

Основные свойства сингулярного разложения:

- $l \times n$-матрица $V = (v_1,\dots,v_n)$ ортогональна, $V^T V = I_n$, столбцы v_j — собственные векторы матрицы FF^T;
- $n \times n$-матрица $U = (u_1,\dots,u_n)$ ортогональна, $U^T U = I_n$, столбцы u_j — собственные векторы матрицы $F^T F$;
- $n \times n$-матрица D диагональна, $D = diag\left(\sqrt{\lambda_1},\dots,\sqrt{\lambda_n}\right)$, $\lambda_j \geq 0$ — собственные значения матриц $F^T F$ и FF^T, $\sqrt{\lambda_j}$ — сингулярные числа матрицы F.

Решение МНК через сингулярное разложение

Псевдообратная F^+, вектор МНК-решения α^*,
МНК-аппроксимация целевого вектора $F\alpha^*$:

$$F^+ = \left(UDV^T VDU^T\right)^{-1} UDV^T = UD^{-1}V^T = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^T;$$

$$\alpha^* = F^+ y = UD^{-1}V^T y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j \left(v_j^T y\right);$$

$$F\alpha^* = P_F y = \left(VDU^T\right)UD^{-1}V^T y = VV^T y = \sum_{j=1}^{n} v_j \left(v_j^T y\right);$$

$$\left\|\alpha^*\right\|^2 = \left\|D^{-1}V^T y\right\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j}\left(v_j^T y\right)^2.$$

Проблема: мультиколлинеарность при $\lambda_j \to 0$.

Проблема мультиколлинеарности и переобучения

Если имеются сингулярные числа, близкие к нулю, то

- матрица $\Sigma = F^T F$ плохо обусловлена;
- решение неустойчивое и неинтерпретируемое, большие коэффициенты $\left\|\alpha^*\right\|$ разных знаков;
- возникает переобучение:
  на обучении $Q(\alpha^*, X^l) = \left\|F\alpha^* - y\right\|^2$ мало;
  на контроле $Q(\alpha^*, X^k) = \left\|F'\alpha^* - y'\right\|^2$ велико.

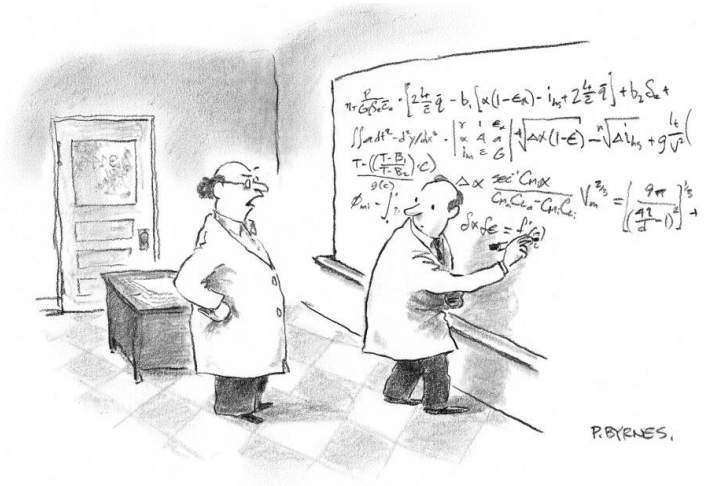
Устранение мультиколлинеарности и переобучения:

- отбор признаков: $f_1,\dots,f_n \to f_{j_1},\dots,f_{j_m}$, $m << n$
- регуляризация: $\|\alpha\| \to \min$;
- преобразование признаков: $f_1,\dots,f_n \to g_1,\dots,g_m$, $m << n$

Резюме

- Задача многомерной линейной регрессии может быть решена через сингулярное разложение;
- Мультиколлинеарность приводит к плохой обусловленности, неустойчивости и переобучению;
- Методы устранения мультиколлинеарности также связаны с сингулярным разложением (гребневая регрессия, метод главных компонент)

Источник: Coursera, К.В. Воронцов



Постановка задачи

Прогнозирование дохода клиента

Схема разработки модели (pipeline)

Подготовка данных → **Создание модели** → **Валидация модели**

- Сборка данных
- Исследование данных
- Предварительная подготовка

- Выбор метода
- Конфигурация модели
- Отбор лучших переменных

- Оценка качества

ДАННЫЕ — ключевая составляющая любой аналитической модели!

Препроцессинг данных

id	AGE	GENDER	TARGET_TOTAL_INCOME
1904	32	1	40000.00
1905	33	2	18000.00
1906	-2019	1	35000.00
1907	22	2	53000.00
1908	24	1	0.00
1909	28	.	70000.00
1910	44	1	300000000000.00
1912	62	2	16000.00
1912	62	2	16000.00
1913	33	1	40000.00
1914	38	1	60000.00
1915	32	2	100000.00
1916	31	2	70000.00
1917	34	1	80000.00
1918	33	2	35000.00
1919	28	2	18000.00

1. Обработка пропущенных значений
2. Определение и обработка выбросов

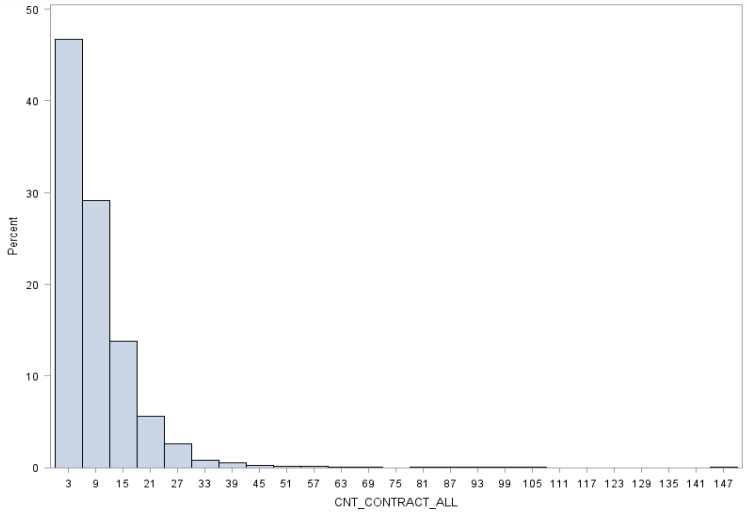
Стратегии

- Сохранить
- Удалить
- Заменить

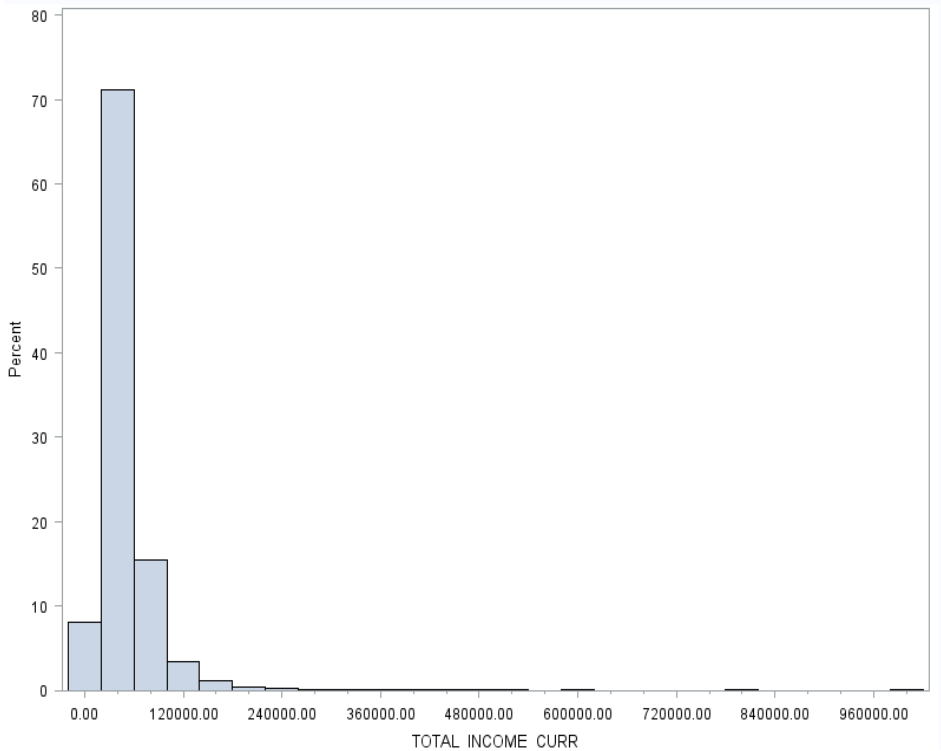
Методы определения выбросов

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
TARGET_TOTAL_INCOME	5000	45962	39402	229811317	0	1000000
GENDER	4977	1.56078	0.49634	7768	1.00000	2.00000
AGE	5000	38.71260	12.01817	193563	18.00000	74.00000
EDUCATION	5000	5.23900	1.57771	26195	1.00000	7.00000
EMAIL_FLAG	5000	0.26500	0.44138	1325	0	1.00000
PRIMARY_INCOME_CURR	5000	40156	36864	200781922	0	1000000
CNT_CONTRACT_ALL	5000	8.31640	8.43069	41582	1.00000	148.00000
MICRO_CREDIT_SUM	976	57148	122191	55776138	435.00000	1531760
LIMIT_SUM	5000	723968	1277711	3619840796	0	32321849
LIMIT_MAX	5000	308581	601953	1542905493	0	21750000
LIMIT_MIN	5000	22917	105931	114582657	0	2886800
LIMIT_CC_AVG	3363	48073	50843	161669861	0	695000
COUNT_DELINQUENCY_30PL	4980	9.43795	16.84773	47001	0	213.00000
PAYMENT_NEXT_MAX	4813	12424	48260	59798226	0	1237804
PAYMENT_NEXT_AVG	4813	6933	22390	33366790	0	625402
COUNT_ACTIVE_CONTRACT	5000	1.84700	1.80395	9235	0	16.00000
REPAYMENT_SUM	4426	437619	840052	1936901216	0	16360000

- Статистические
- Гистограмма
- Расчет станд. отклонения (Z-score)



TARGET_TOTAL_INCOME



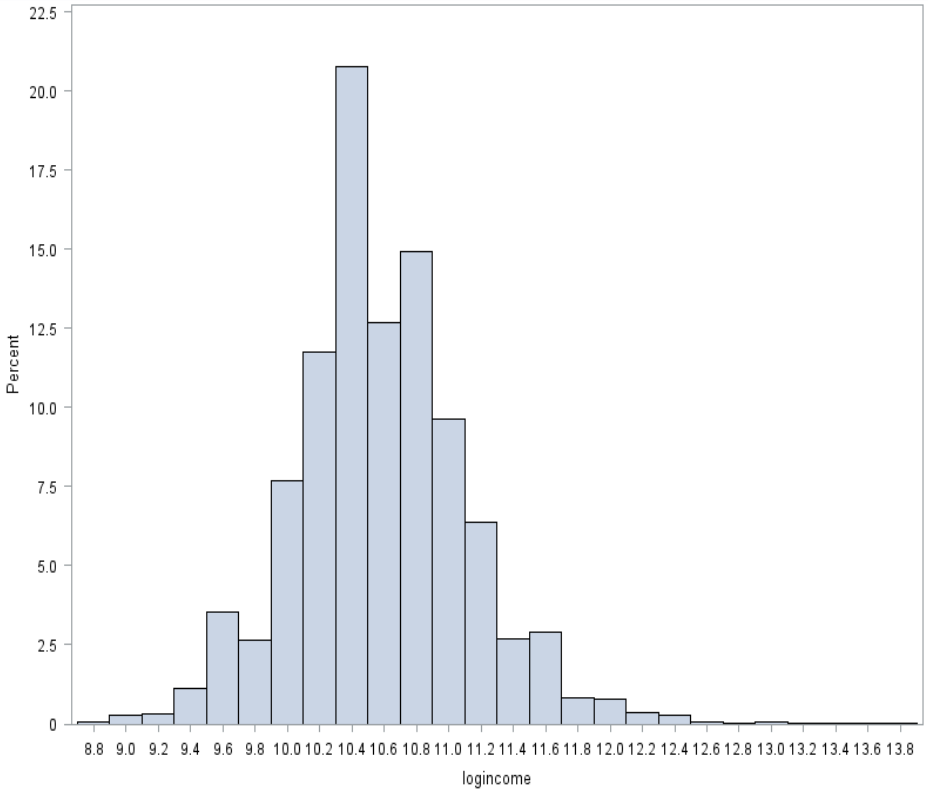
- Форма распределения отлична от нормальной
- Наличие выбросов

Moments			
N	5000	Sum Weights	5000
Mean	45962.2634	Sum Observations	229811317
Std Deviation	39401.6684	Variance	1552491475
Skewness	9.86994354	Kurtosis	185.924229
Uncorrected SS	1.83236E13	Corrected SS	7.7609E12
Coeff Variation	85.7261273	Std Error Mean	557.223739

Basic Statistical Measures			
Location		Variability	
Mean	45962.26	Std Deviation	39402
Median	38000.00	Variance	1552491475
Mode	30000.00	Range	1000000
		Interquartile Range	24450

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000000
99%	170000
95%	100000
90%	79950
75% Q3	52000
50% Median	38000
25% Q1	27550
10%	20000
5%	16000
1%	12000
0% Min	0

LOG трансформация



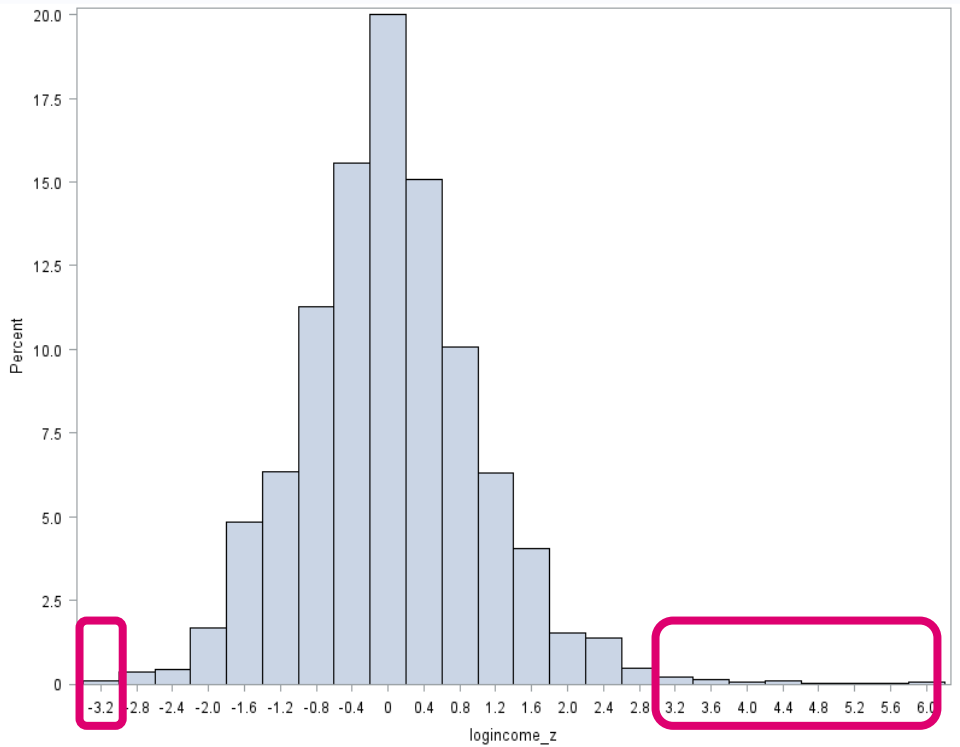
Любая трансформация данных делает вашу модель **менее интерпретируемой**

Moments			
N	4997	Sum Weights	4997
Mean	10.5661338	Sum Observations	52798.9706
Std Deviation	0.54678528	Variance	0.29897415
Skewness	0.46875384	Kurtosis	1.43665905
Uncorrected SS	559374.664	Corrected SS	1493.67483
Coeff Variation	5.17488509	Std Error Mean	0.00773503

Basic Statistical Measures			
Location		Variability	
Mean	10.56613	Std Deviation	0.54679
Median	10.54534	Variance	0.29897
Mode	10.30895	Range	5.00565
		Interquartile Range	0.61904

Quantiles (Definition 5)	
Level	Quantile
100% Max	13.81551
99%	12.04355
95%	11.51293
90%	11.28978
75% Q3	10.85900
50% Median	10.54534
25% Q1	10.23996
10%	9.90349
5%	9.68034
1%	9.39266
0% Min	8.80986

Z-score



Значения отклоняющиеся от среднего более чем на +/- 3*Sigma удаляются из обучающей выборки

Moments			
N	4997	Sum Weights	4997
Mean	0	Sum Observations	0
Std Deviation	1	Variance	1
Skewness	0.46875384	Kurtosis	1.43665905
Uncorrected SS	4996	Corrected SS	4996
Coeff Variation	.	Std Error Mean	0.01414638

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	1.00000
Median	-0.03803	Variance	1.00000
Mode	-0.47035	Range	9.15469
		Interquartile Range	1.13214

Quantiles (Definition 5)	
Level	Quantile
100% Max	5.9426924
99%	2.7020111
95%	1.7315602
90%	1.3234594
75% Q3	0.5356128
50% Median	-0.0380266
25% Q1	-0.5965304
10%	-1.2118948
5%	-1.6199957
1%	-2.1461292
0% Min	-3.2119939

Построение модели

The CORR Procedure

4 Variables:	TARGET_TOTAL_INCOME	LIMIT_SUM	LIMIT_MAX	LIMIT_MIN

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
TARGET_TOTAL_INCOME	5000	42714	49357	213571256	0	2500036	TOTAL_INCOME_CURR
LIMIT_SUM	3021	803236	1471075	2426577231	0	36734657	
LIMIT_MAX	3021	343007	758892	1036222659	0	20000000	
LIMIT_MIN	3021	21818	92138	65913193	0	3187500	

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	TARGET_TOTAL_INCOME	LIMIT_SUM	LIMIT_MAX	LIMIT_MIN
TARGET_TOTAL_INCOME	1.00000	0.22889	0.19330	0.02565
		<.0001	<.0001	0.1587
TOTAL_INCOME_CURR	5000	3021	3021	3021
LIMIT_SUM	0.22889	1.00000	0.88834	0.07931
	<.0001		<.0001	<.0001
	3021	3021	3021	3021
LIMIT_MAX	0.19330	0.88834	1.00000	0.13256
	<.0001	<.0001		<.0001
	3021	3021	3021	3021
LIMIT_MIN	0.02565	0.07931	0.13256	1.00000
	0.1587	<.0001	<.0001	
	3021	3021	3021	3021

Page Break

Переменные с сильной корреляцией должны быть исключены из модели

Самостоятельная работа

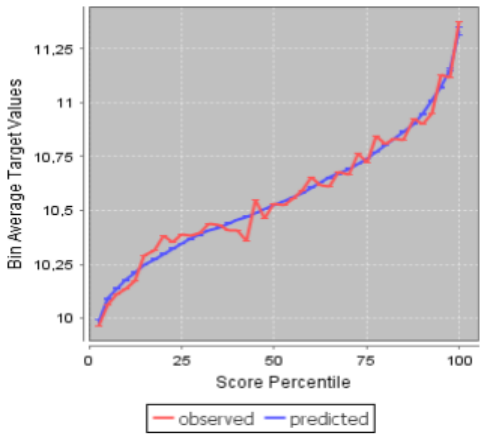
Пример модели

Scorecard Model Editor: Variables - Model_inc_V2.mb

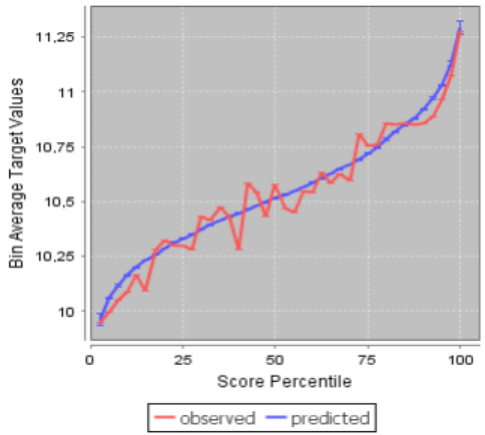
	Initial	In	Tier	Binning Nam... Bin Label	Variable Description	Constraint	Marginal Contribution (Ver. 4) Training	Test	Training
⊞	☐	☑	1	LIMIT_SUM			0,074	0,072	
⊞	☐	☑	1	GENDER			0,035	0,045	
⊞	☐	☑	1	EDUCATION			0,029	0,033	
⊞	☐	☑	1	AGE			0,028	0,040	
⊞	☐	☑	1	LIMIT_MIN			0,024	0,018	
⊞	☐	☑	1	LIMIT_CC_AVG			0,019	0,013	
⊞	☐	☑	1	PAYMENT_NEXT_MAX			0,013	0,009	

Score Bin Plot & Table

Model_inc_V2_score, setID: train



Model_inc_V2_score, setID: test



Исключаем:

1. PRIMARY_INCOME_CURR (целевая)
2. LIMIT_MAX (корреляция)
3. PAYMENT_NEXT_AVG (корреляция)
4. DATE (дата)

Добавляем:

1. REPAYMENT2LIMIT (отношение)

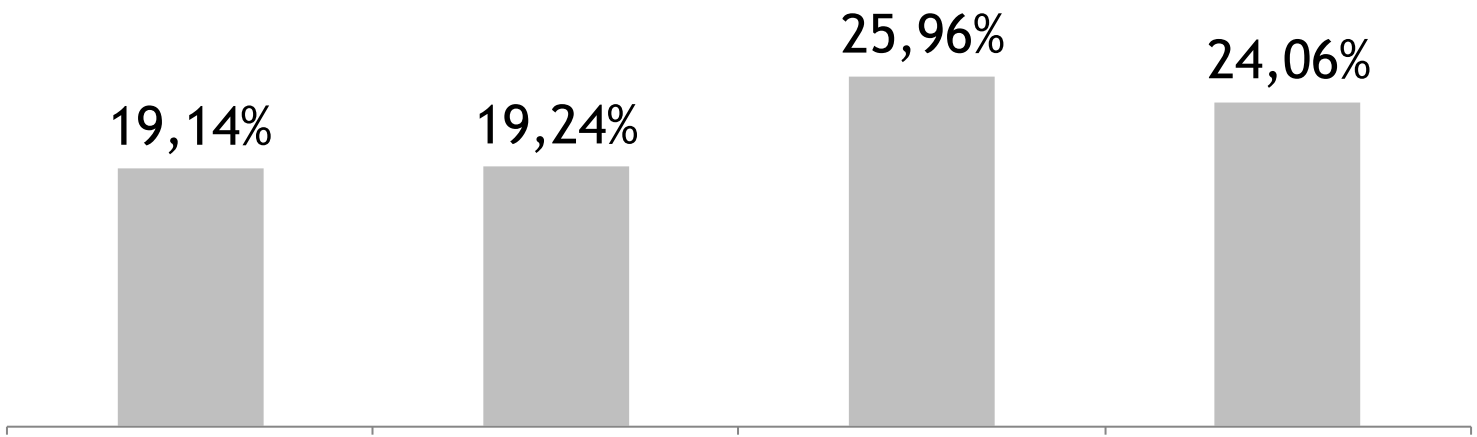
Summary Score Statistics

setID	Score Variable	R-Squared	RMSE	NSSE
train	Model_inc_V2_score	0,333	0,424	0,667
test	Model_inc_V2_score	0,321	0,433	0,679

Матрица корреляций

	TARGET_TOTAL_INCOME	GENDER	AGE	EDUCATION	EMAIL_FLAG	PRIMARY_INCOME_CURR	CNT_CONTRACT_ALL	MICRO_CREDIT_SUM	LIMIT_SUM	LIMIT_MAX	LIMIT_MIN	LIMIT_CC_AVG	COUNT_DELINQUENCY_30PL	PAYMENT_NEXT_MAX	PAYMENT_NEXT_AVG	COUNT_ACTIVE_CONTRACT	REPAYMENT_SUM	PAYMENT2LIMIT
TARGET_TOTAL_INCOME	1,00	0,13	0,06	0,19	0,10	0,95	0,07	0,14	0,34	0,29	0,04	0,28	0,08	0,09	0,08	0,14	0,34	0,09
GENDER	0,13	1,00	0,17	0,06	0,02	0,15	0,07	0,09	0,05	0,06	0,00	0,06	0,03	0,03	0,03	0,07	0,05	0,01
AGE	0,06	0,17	1,00	0,08	0,14	0,10	0,19	0,12	0,10	0,04	0,04	0,06	0,07	0,02	0,01	0,07	0,11	0,10
EDUCATION	0,19	0,06	0,08	1,00	0,05	0,19	0,02	0,08	0,17	0,15	0,03	0,16	0,01	0,04	0,03	0,05	0,14	0,03
EMAIL_FLAG	0,10	0,02	0,14	0,05	1,00	0,10	0,04	0,08	0,05	0,05	0,00	0,06	0,06	0,04	0,03	0,10	0,04	0,07
PRIMARY_INCOME_CURR	0,95	0,15	0,10	0,19	0,10	1,00	0,04	0,10	0,32	0,27	0,05	0,25	0,07	0,08	0,08	0,10	0,32	0,07
CNT_CONTRACT_ALL	0,07	0,07	0,19	0,02	0,04	0,04	1,00	0,60	0,38	0,17	0,12	0,06	0,38	0,09	0,03	0,54	0,38	0,01
MICRO_CREDIT_SUM	0,14	0,09	0,12	0,08	0,08	0,10	0,60	1,00	0,24	0,13	0,12	0,06	0,22	0,13	0,11	0,19	0,24	0,04
LIMIT_SUM	0,34	0,05	0,10	0,17	0,05	0,32	0,38	0,24	1,00	0,86	0,07	0,38	0,18	0,23	0,18	0,36	0,86	0,13
LIMIT_MAX	0,29	0,06	0,04	0,15	0,05	0,27	0,17	0,13	0,86	1,00	0,16	0,33	0,08	0,19	0,18	0,22	0,60	0,21
LIMIT_MIN	0,04	0,00	0,04	0,03	0,00	0,05	0,12	0,12	0,07	0,16	1,00	0,22	0,06	0,01	0,05	0,07	0,07	0,04
LIMIT_CC_AVG	0,28	0,06	0,06	0,16	0,06	0,25	0,06	0,06	0,38	0,33	0,22	1,00	0,00	0,09	0,10	0,17	0,27	0,15
COUNT_DELINQUENCY_30PL	0,08	0,03	0,07	0,01	0,06	0,07	0,38	0,22	0,18	0,08	0,06	0,00	1,00	0,11	0,08	0,24	0,16	0,00
PAYMENT_NEXT_MAX	0,09	0,03	0,02	0,04	0,04	0,08	0,09	0,13	0,23	0,19	0,01	0,09	0,11	1,00	0,88	0,07	0,18	0,10
PAYMENT_NEXT_AVG	0,08	0,03	0,01	0,03	0,03	0,08	0,03	0,11	0,18	0,18	0,05	0,10	0,08	0,88	1,00	0,02	0,13	0,12
COUNT_ACTIVE_CONTRACT	0,14	0,07	0,07	0,05	0,10	0,10	0,54	0,19	0,36	0,22	0,07	0,17	0,24	0,07	0,02	1,00	0,23	0,45
REPAYMENT_SUM	0,34	0,05	0,11	0,14	0,04	0,32	0,38	0,24	0,86	0,60	0,07	0,27	0,16	0,18	0,13	0,23	1,00	0,12
PAYMENT2LIMIT	0,09	0,01	0,10	0,03	0,07	0,07	0,01	0,04	0,13	0,21	0,04	0,15	0,00	0,10	0,12	0,45	0,12	1

R^2 (test)



del DATE del ID del missing GENDER MVI	del DATE del ID del missing GENDER MVI del correlated	del DATE del ID del missing GENDER MVI del correlated del outliers of TARGET	del DATE del ID del missing GENDER MVI del correlated del outliers of TARGET del all +/- 3*Sigma

APPENDIX

Как бороться с мультиколлинеарностью?

Гребневая регрессия (Ridge Regression)

Штраф за увеличение нормы вектора весов $\|\alpha\|$:

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \frac{\tau}{2}\|\alpha\|^2,$$

где τ — неотрицательный параметр решуляризации.

Модифицированное МНК-решение (τI_n — "гребень")

$$\alpha_\tau^* = \left(F^T F + \tau I_n\right)^{-1} F^T y.$$

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения α_τ^* и МНК-аппроксимация целевого вектора $F\alpha_\tau^*$:

$$\alpha_\tau^* = U\left(D^2 + \tau I_n\right)^{-1} DV^T y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j \left(v_j^T y\right);$$

$$F\alpha_\tau^* = VDU^T \alpha_\tau^* = V diag\left(\frac{\sqrt{\lambda_j}}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j \left(v_j^T y\right);$$

$$\left\|\alpha_\tau^*\right\|^2 = \left\|\left(D^2 + \tau I_n\right)^{-1} DV^T y\right\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{\left(\lambda_j + \tau\right)^2} \left(v_j^T y\right)^2$$

$F\alpha_\tau^* \neq F\alpha^*$, но зато решение становится гораздо устойчивее.

Выбор параметра регуляризации

Контрольная выборка: $X^k = \left(x_i', y_i'\right)_{i=1}^k$:

$$\underset{k\times n}{F'} = \begin{pmatrix} f_1\left(x_1'\right) & \ldots & f_n\left(x_1'\right) \\ \ldots & \ldots & \ldots \\ f_1\left(x_k'\right) & \ldots & f_n\left(x_k'\right) \end{pmatrix}, \quad \underset{k\times 1}{y'} = \begin{pmatrix} y_1' \\ \ldots \\ y_k' \end{pmatrix}.$$

Вычисление функционала Q на контрольных данных T раз потребует $O\left(kn^2 + knT\right)$ операций:

$$Q(\tau) = \left\|F'\alpha_\tau^*\right\|^2 = \left\| \underset{k\times n}{F'U}\, diag\left(\frac{\sqrt{\lambda_j}}{\lambda_j + \tau}\right) \underset{k\times n}{V^T} y - y' \right\|^2.$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Регуляризация сокращает "эффективную размерность"

Сжатие (shrinkage) или сокращение весов (weight decay):

$$\left\|\alpha_\tau^*\right\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{\left(\lambda_j + \tau\right)^2} \left(v_j^T y\right)^2 < \left\|\alpha^*\right\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} \left(v_j^T y\right)^2$$

Роль размерности играет след проекционной матрицы:

$$tr\, F\left(F^T F\right)^{-1} F^T = tr\left(F^T F\right)^{-1} F^T F = tr\, I_n = n.$$

При использовании решуляризации:

$$tr\, F\left(F^T F + \tau I_n\right)^{-1} F^T = tr\, diag\left(\frac{\lambda_j}{\lambda_j + \tau}\right) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} < n$$

Источник: Coursera, К.В. Воронцов

Как бороться с мультиколлинеарностью?

Регрессия LASSO

LASSO — Least Absolute Shrinkage and Selection Operator, два эквивалентных варианта постановки задаци:

$$Q(\alpha) = \|F\alpha - y\|^2 \to \min_{\alpha} \quad \text{при} \quad \sum_{j=1}^{n} |\alpha_j| \le \chi;$$

$$Q(\alpha) = \|F\alpha - y\|^2 + \tau \sum_{j=1}^{n} |\alpha_j| \to \min_{\alpha};$$

После замены переменных

$$\begin{cases} \alpha_j = \alpha_j^+ - \alpha_j^-; \\ |\alpha_j| = \alpha_j^+ - \alpha_j^-; \end{cases} \quad \alpha_j^+ \ge 0; \quad \alpha_j^- \ge 0.$$

ограничения принимают канонический вид:

$$\sum_{j=1}^{n} \alpha_j^+ + \alpha_j^- \le \chi; \quad \alpha_j^+ \ge 0; \quad \alpha_j^- \ge 0.$$

Чем меньше χ, тем больше j таких, что $\alpha_j^+ = \alpha_j^- = 0$.

Резюме:

LASSO обнуляет веса и приводи к отбору признаков в линейных моделях.

Источник: Coursera, К.В. Воронцов

Как бороться с мультиколлинеарностью?

Метод главных компонент (PCA)

Постановка задачи PCA (principal component analysis):

$f_1(x),\dots, f_n(x)$ — исходные числовые признаки

$g_1(x),\dots, g_m(x)$ — новые числовые признаки, $m \le n$;

Требование: старые признаки должны линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \quad j = 1,\dots,n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке $x_1,\dots,x_l$:

$$\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 \to \min_{\{g_s(x_i)\},\{u_{js}\}}$$

Матричные обозначения

Матрицы "объекты-признаки", старая и новая

$$\underset{l\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix}; \quad \underset{l\times m}{G} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_l) & \dots & g_m(x_l) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \quad \hat{F} = GU^T \overset{\text{хотим}}{\approx} F.$$

Найти: и новые признаки G, и преобразование U

$$\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 = \left\|GU^T - F\right\|^2 \to \min_{G,U}$$

Основная теорема метода PCA

Если $m \le rk\ F$, то минимум $\left\|GU^T - F\right\|^2$ достигается, когда столбцы U — это собственные векторы матрицы F^TF, соответствующие m максимальным собственным значениям $\lambda_1,\dots,\lambda_m$, а матрица $G - FU$.

При этом:

- матрица U ортонормированна: $U^TU = I_m$;
- матрица G ортогональна: $G^TG = \Lambda = diag(\lambda_1,\dots,\lambda_m$;
- $U\Lambda = F^TFU$; $G\Lambda = FF^TG$;
- $\left\|GU^T - F\right\|^2 = \|F\|^2 - tr\ \Lambda = \sum_{j=m+1}^{n}\lambda_j$.

Связь с сингулярным разложением

Если взять $m = n$, то:

- $\left\|GU^T - F\right\|^2 = 0$;
- представление $\hat{F} = GU^T = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$: $F = GU^T = V\sqrt{\Lambda}U^T$; $U^TU = I_m$; $V^TV = I_m$.
- линейное преобразование U работает в обе стороны: $F = GU^T$; $G = FU$.

Поскольку новые признаки некоррелированы $\left(G^TG = \Lambda\right)$, преобразование U называется декоррелирующим (или преобразованием Карунена-Лоэва).

Эффективная размерность выборки

Упорядочим с.з. F^TF по убыванию: $\lambda_1 \ge \dots \ge \lambda_n \ge 0$.

Эффективная размерность выборки — это наименьшее целое m, при котором:

$$E_m = \frac{\left\|GU^T - F\right\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \le \varepsilon$$

Решение задачи НК для МЛР в новых признаках

Задача наименьших квадратов для МЛР: $\|F\alpha - y\|^2 \to \min_{\alpha}$.

Заменим $\underset{l\bullet n}{F}$ на ее приближение $\underset{l\bullet m}{G} \bullet \underset{m\bullet n}{U^T}$, предполагая $m \le n$:

$$\left\|GU^T\alpha - y\right\|^2 = \|G\beta - y\|^2 \to \min_{\beta}$$

Связь нового и старого вектора коэффициентов:

$$\beta = U^T\alpha; \quad \alpha = U\beta.$$

Решение задачи наименьших квадратов относительно β (единственное отличие — m слагаемых вместо n):

$$\beta^* = D^{-1}V^Ty; \quad \alpha^* = UD^{-1}V^Ty = \sum_{j=1}^{m}\frac{1}{\sqrt{\lambda_j}}u_j\left(v_j^Ty\right);$$

$$G\beta^* = VV^Ty = \sum_{j=1}^{m}v_j\left(v_j^Ty\right)$$

Резюме:

- Метод главных компонент позволяет приближать матрицу ее низкоранговым разложением;
- Для этого достаточно взять из SVD-разложения первые m сингулярных чисел и векторов матрицы;
- Этот прием широко используется в анализе данных — в задачах регрессии, классификации, сжатия данных и др.

Источник: Coursera, К.В. Воронцов

Обучение

Курсы на coursera.org

Введение в машинное обучение
Воронцов К.В., Соколов Е.
National Research University Higher School of Economics, Yandex School of Data Analysis

Machine Learning
Andrew Ng
Stanford University

Neural Networks and Deep Learning
Andrew Ng
deeplearning.ai

Improving Deep Neural Networks
Andrew Ng
deeplearning.ai

Structuring Machine Learning Projects
Andrew Ng
deeplearning.ai

Convolutional Neural Networks
Andrew Ng
deeplearning.ai

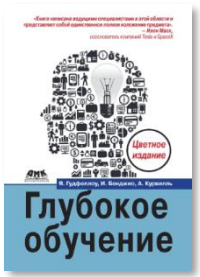
Sequence Models
Andrew Ng
deeplearning.ai

Книги на русском языке

Введение в машинное обучение с помощью Python

Мюллер А., Гидо С.

O'Reilly Media, 2017. — 392 с.



Глубокое обучение

Ян Гудфеллоу, Иошуа Бенджио, Аарон Курвилль

ДМК-Пресс, 2018. — 652 с.

Глубокое обучение: погружение в мир нейронных сетей

Николенко С., Кадурин А., Архангельская Е.

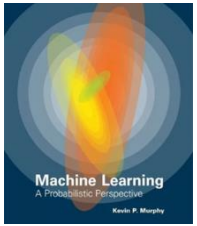
СПб.: Питер, 2018. — 480 с.

Библиотека Keras — инструмент глубокого обучения

Антонио Джулли, Суджит Пал

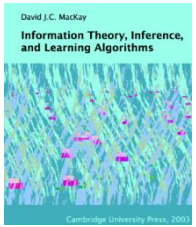
ДМК-Пресс, 2017. — 296 с.

Книги на английском языке



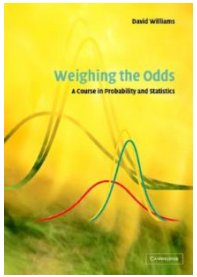
Machine Learning: A Probabilistic Perspective
Murphy K.P.

Massachusetts Institute of Technology, 2012. — 1067 p.



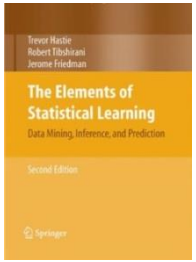
Information Theory, Inference, and Learning Algorithms
MacKay D.J.C.

Cambridge University Press, 2003. — 640 p



Weighing the Odds: A Course in Probability and Statistics
David Williams

Cambridge University Press, 2001. — 568 p.



The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Hastie T., Tibshirani R., Friedman J.

Second Edition (corrected 5th printing). — Springer, 2009. — 763 p.

Лекции и полезные ресурсы

https://www.youtube.com/playlist?list=PLJOzdkh8T5kp99tGTEFjH_b9zqEQiiBtC

Лекции ШАД от Константина Воронцова (д.ф.м.н. МФТИ, ВШЭ)

https://www.lektorium.tv/speaker/2691

Лекции Сергея Николенко (автор книги «Глубокое обучение: погружение в мир нейронных сетей»)

http://efimov-ml.com/

Сайт с лекциями Дмитрия Ефимова, а также с Python-скриптами (Jupiter Notebooks), презентациями и ссылками

http://deepbayes.ru/2017/

Лекции ВШЭ по Байесовским методам в глубинном обучении. Лекторы: Дмитрий Ветров, Дмитрий Кропотов, Евгений Соколов, Сергей Бартунов, Арсений Ашуха и др.

https://arxiv.org/

Ресурс с самыми свежими препринтами научных статей по ML/AI/DS (см. раздел Computer Science)

http://scikit-learn.org/
http://www.numpy.org/
https://pandas.pydata.org/
https://keras.io/
http://devdocs.io/tensorflow/

Подробные мануалы популярных библиотек в Python

https://stackoverflow.com/

Ресурс для обсуждения практических вопросов по программированию на Python и других языках

http://ods.ai/

Русскоязычное сообщество датасайнтистов