Федеральное государственное бюджетное образовательное учреждение
высшего образования
Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

УТВЕРЖДАЮ
декан факультета вычислительной
математики и кибернетики

_____ /И.А. Соколов /
«27» сентября 2022г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине

Прикладные задачи анализа данных

Уровень высшего образования:

бакалавриат

Направление подготовки / специальность:

01.03.02 "Прикладная математика и информатика" (3++)

Направленность (профиль) ОПОП:

Искусственный интеллект и анализ данных

Форма обучения:

очная

Рассмотрен и утвержден

на заседании Ученого совета факультета ВМК

(протокол №7, от 27 сентября 2022 года)

Москва 2022

1. ФОРМЫ И ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

В процессе и по завершении изучения дисциплины оценивается формирование у студентов следующих компетенций:

Планируемые результаты обучения по дисциплине (модулю)		
Содержание и код компетенции.	**Индикатор (показатель) достижения компетенции**	**Планируемые результаты обучения по дисциплине, сопряженные с индикаторами достижения компетенций**
ПК-11. Способен анализировать, разрабатывать, внедрять и выполнять организационно-технические и экономические процессы с применением технологий и систем искусственного интеллекта	ПК-11.1. Использует знание рынка информационных систем и информационно-коммуникационных технологий, методов математического моделирования и искусственного интеллекта для анализа и разработки организационно-технических и экономических процессов ПК-11.2. Решает задачи по построению организационно-технических и экономических процессов с применением информационных технологий и систем искусственного интеллекта	Знать: основные принципы решения задач анализа данных и построения алгебраических выражений над алгоритмами для решения таких за-дач Уметь решать современные прикладные задачи анализа данных: классификацию (распознавание образов), регрессию, прогнозирование, кластеризацию, строить ансамбли над алгоритмами Владеть современными программными пакетами анализа данных, навыками написания отчётов и подготовки докладов о решении задачи

1.1. Текущий контроль успеваемости

Текущий контроль успеваемости осуществляется путем оценки результатов выполнения заданий практических (семинарских) занятий, самостоятельной работы, предусмотренных учебным планом и посещения занятий/активность на занятиях.

В качестве оценочных средств текущего контроля успеваемости предусмотрены:

выполнение заданий на практических (семинарских) занятиях

Примеры заданий для практических (семинарских) занятий

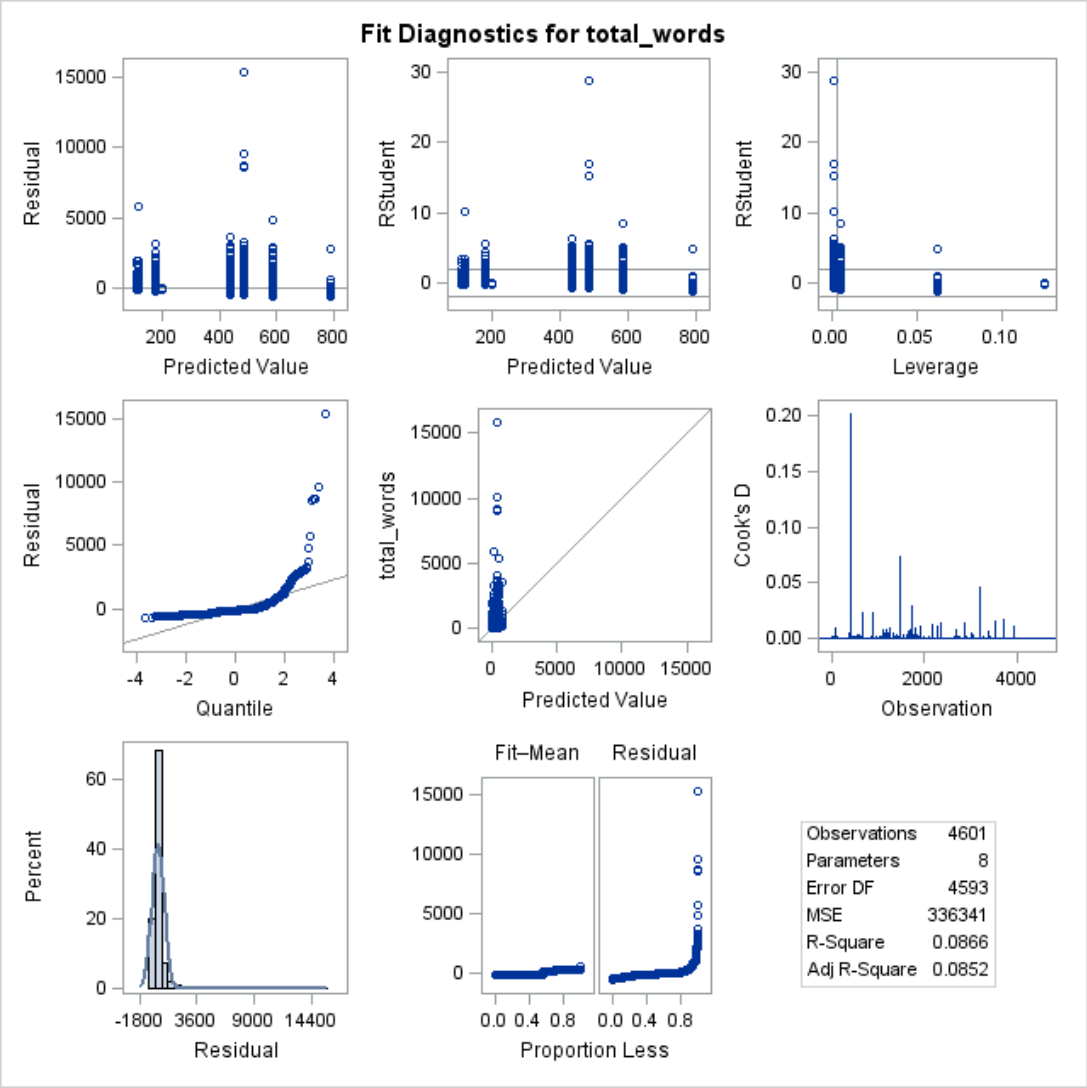
1. Дисперсионный анализ

Используя набор данных text_data написать программу для проверки предположения, что размер текста в файле (total_words) зависит от авторства и признака, является ли письмо выжным, а также выяснить какие авторы пишут тексты примерно одинакового объема, а какие нет.
Предположим, что Вы получили частичный вывод программы, представленный ниже.

Dependent Variable: total_words

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	7	146409864	20915695	62.19	<.0001
Error	4593	1544815629	336341		
Corrected Total	4600	1691225494			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Author	3	29483945.4	9827981.8	29.22	<.0001
importance	1	115367988.8	115367988.8	343.01	<.0001
Author*importance	3	1557930.2	519310.1	1.54	0.2010



Level of Author	N	total_words	
		Mean	Std Dev
Ivanov	**735**	177.503401	333.115084
Petrov	**203**	601.256158	836.150594
Sidorov	**2439**	296.598196	692.363455
Smirnov	**1224**	267.558007	470.180091

Least Squares Means

Author	total_words LSMEAN	LSMEAN Number
Ivanov	349.046771	1
Petrov	748.973831	2
Sidorov	300.696418	3
Smirnov	271.569527	4

i/j	1	2	3	4
1	_	<.0001	0.0633	0.0067
2	<.0001	_	<.0001	<.0001
3	0.0633	<.0001	_	0.1518
4	0.0067	<.0001	0.1518	_

Ответьте на следующие вопросы (везде считать, что уровень значимости равен 0.01):

1) Принята ли базовая гипотеза дисперсионного анализа?
2) Есть ли выбросы в наборе данных
3) Какие предположения дисперсионного анализа нарушены (если нарушены) в данной задаче?
4) Нужно ли использовать в модели переменную Author? Переменную Importance? Их
5) Какие пары авторов неразличимы с точки зрения описания вариации переменной total_words?

2. Регрессионный анализ.

Предположим, что изначально в наборе данных примеров не важных писем было очень много – 99% от выборки. Далее с помощью подхода oversampling и процедуры surveyselect выборка была сбалансирована, т.е. получен набор balanced_text_data, где пропорция важных и обычных текстов уже 1:1. На наборе данных balanced_text_data постройте и сохраните модель на основе логистической регрессии для прогнозирования признака, является ли текст важным. При этом должен быть осуществлен отбор значимых переменных комбинированным пошаговым (stepwise) методом. Порог уровня значимости при добавлении переменной должен быть 0.01, а при удалении 0.05. Должны быть выведены ROC кривые для каждого шага. Напишите программу, которая применит полученную модель к набору данных той же структуры с именем score_text_data, где в переменной p_importance будет записана корректная с учетом балансировки тренировочного набора вероятность того, что текст является важным.
Предположим, что Вы получили частичный вывод программы, представленный ниже. Ответьте на следующте вопросы (везде считать, что уровень значимости равен 0.01):

1) Принята ли базовая гипотеза регрессионного анализа?

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	3995.6243	27	<.0001
Score	2487.9121	27	<.0001
Wald	808.9079	27	<.0001

2) Какие из перечисленных переменных можно исключить из модели без существенной потери качества? Если их несколько, то можно ли их исключить все?

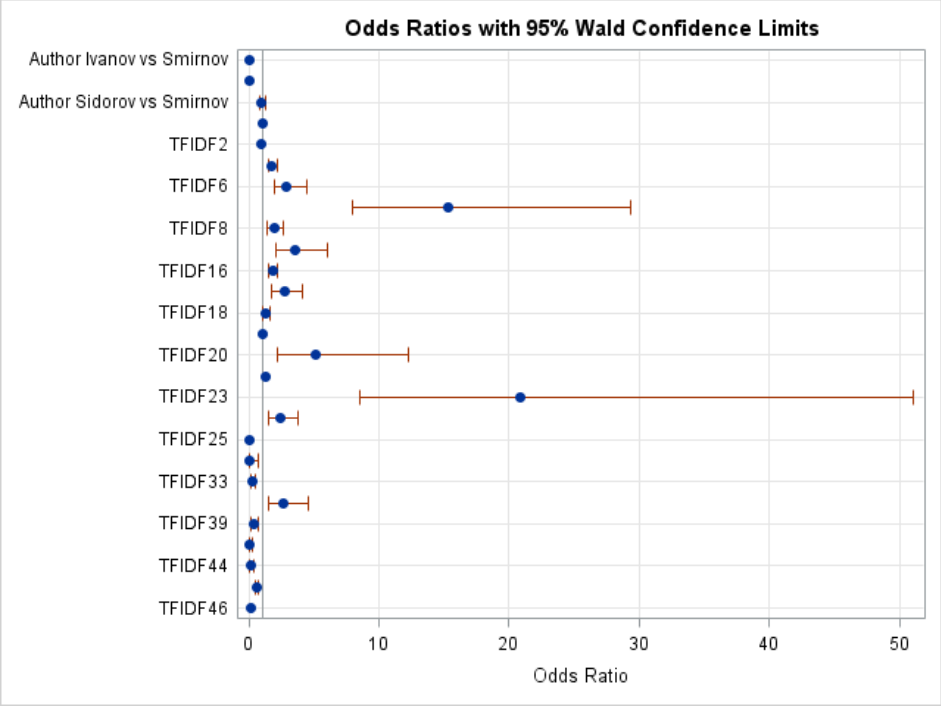
Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard**	**Wald**	**Pr > ChiSq**
Intercept		1	-3.0090	0.1984	229.9632	<.0001
Author	Ivanov	1	-2.6571	0.3837	47.9466	<.0001
Author	Petrov	1	-0.8639	0.3381	6.5296	0.0106
Author	Sidorov	1	1.7507	0.1733	102.0490	<.0001
total_words		1	0.00178	0.000187	90.1722	<.0001
TFIDF18		1	0.2710	0.1077	6.3300	0.0119

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard**	**Wald**	**Pr > ChiSq**
TFIDF19		1	0.0797	0.0307	6.7381	0.0094
TFIDF20		1	1.6338	0.4437	13.5609	0.0002
TFIDF21		1	0.2166	0.0430	25.3127	<.0001
TFIDF29		1	-3.6589	1.7155	4.5493	0.0329
TFIDF33		1	-1.3133	0.3384	15.0642	0.0001
TFIDF36		1	0.9555	0.2918	10.7218	0.0011

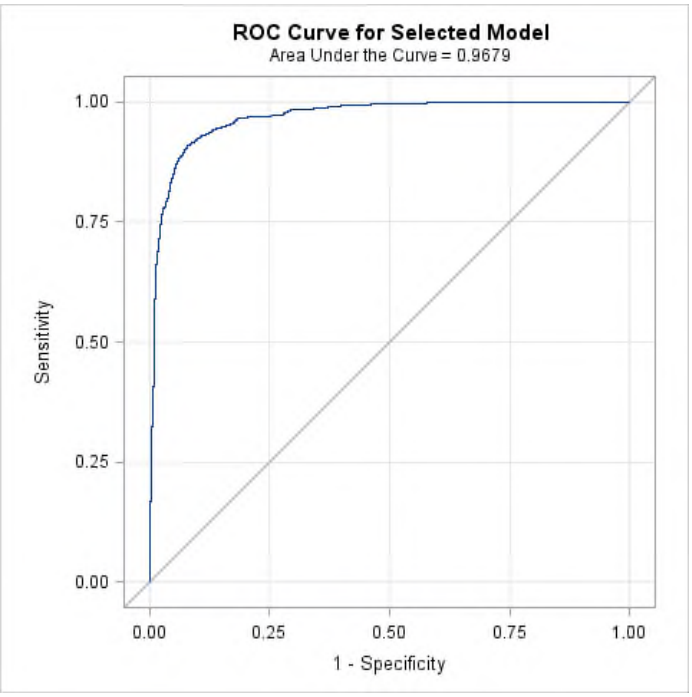
3) Как видно из таблицы ниже процесс отбора переменных остановился на 27 шаге. Почему?

ummary of Stepwise Selection								
Step	**Effect**		**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**	**Variable Label**
	Entered	**Removed**						
1	**TFIDF21**		1	1	675.7404		<.0001	
2	**Author**		3	2	546.3350		<.0001	
3	**TFIDF23**		1	3	290.8007		<.0001	
4	**TFIDF7**		1	4	289.8012		<.0001	
5	**TFIDF16**		1	5	201.8526		<.0001	
	...	**...**	...	...	...	...	...	...
25	**TFIDF18**		1	25	6.7127		0.0096	
26	**TFIDF47**		1	26	6.6308		0.0100	
27		**TFIDF47**	1	25		3.5285	0.0603	

4) Какая из переменных оказывает наибольшее влияние на отклик? При всех остальных равных переменных, если автором текста является Иванов, то для его текста вероятность высокой важности ниже чем у Сидорова или выше?



5) Примерно каким будет уровень ложно положительных срабатываний если выбрать порог таким, чтобы не пропустить ни одного важного сообщения? При каком значении ошибки перового и второго рода будут совпадать?



1.2. Промежуточная аттестация

Промежуточная аттестация осуществляется в форме экзамена
В качестве средств, используемых на промежуточной аттестации предусматривается:
Билеты

1.3. Типовые задания для проведения промежуточной аттестации

Вопросы к экзамену

1. Основные принципы работы шага обработки данных.
2. Работа со структурированными наборами данных и массивами.
3. Процедуры проверки гипотез и дисперсионного анализа.
4. Процедуры построения линейных регрессионных моделей. Смешанные линейные регрессионные модели.
5. Проблема мультиколлинеарности, пошаговый отбор переменных, регуляризация, преобразования пространства признаков.
6. Процедуры поиска главных компонент и кластеризации переменных.
7. Процедуры и инструменты для поиска выбросов.
8. Процедуры построения нелинейных регрессий.
9. Анализ таблиц сопряженности, логистическая регрессия.
10. Обобщенные линейные модели, пуассоновская и гамма регрессии.
11. Сравнение и оценка моделей на тестовом наборе данных.

Пример экзаменационного билета

1. Работа со структурированными наборами данных и массивами.
2. Процедуры построения нелинейных регрессий.

2. КРИТЕРИИ ОЦЕНКИ ПО ДИСЦИПЛИНЕ

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	**2 (не зачтено)**	**3 (зачтено)**	**4 (зачтено)**	**5 (зачтено)**
Знания **(виды оценочных средств: приведены в п. 1.2.)**	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения **(виды оценочных средств: приведены в п. 1.2.)**	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки **(владения, опыт деятельности)** **(виды оценочных средств: приведены в п. 1.2..)**	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач