



Machine Learning. Professional Защита проекта

Меня хорошо видно && слышно?



Защита проекта

Тема: Анализ результатов IELTS методами ML



Золотарева Наталья

Цель проекта



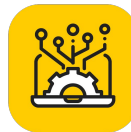
Создание модели способной прогнозировать оценку письменной части языкового экзамена IELTS на основе текста эссе

План работы



1. Исследование данных
2. Генерирование дополнительных признаков
3. Построение моделей
4. Исследовать эффективности сгенерированных признаков

Используемые технологии



1. `spacy` для лемматизации и токенизации
2. TF-IDF для векторизации (при построении моделей)
3. предобученный Word2Vec из Gensim для векторизации (для генерации признаков)
4. `readability` (для генерации признаков)

Данные

Начальные данные:

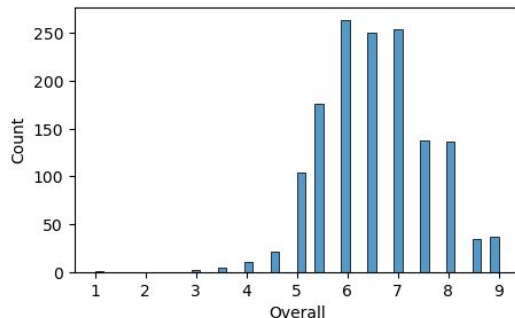
- 1435 ненулевых строк
- КОЛОНКИ

Task_Type - тип задания (описание диаграммы или эссе)

Question - текст вопроса

Essay - текст эссе

Overall - итоговая оценка



Overall

1.0	1
3.0	2
3.5	4
4.0	8
4.5	21
5.0	92
5.5	161
6.0	238
6.5	214
7.0	222
7.5	118
8.0	126
8.5	32
9.0	36



Новые признаки

Word_counter

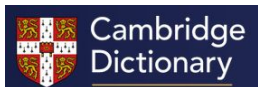
Количество слов в эссе из списков слов разных уровней

A1, A2, B1, B2, C1, C2, AC

<https://dictionary.cambridge.org/dictionary/english/>

Academic Word List (AWL)

https://simple.wiktionary.org/wiki/Wiktionary:Academic_word_list



Similarity

cosine_similarity между эмбедингом вопроса и суммарным эмбедингом кадого абзаца, полученного суммой эмбедингов их предложений

Для получения эмбедингов использован предобученный word2vec 'GoogleNews-vectors-negative300.bin.gz'

readability

Реализация метрик удобочитаемости, основанных на поверхностных характеристиках (линейные регрессии, основанные на количестве слов, слогов и предложений).

- $$4.71 \left(\frac{\text{characters}}{\text{words}} \right) + 0.5 \left(\frac{\text{words}}{\text{sentences}} \right) - 21.43$$

- **characters per word**

- **LIX (швед. Läsbarhetsindex) — индекс удобочитаемости**

$$\text{LIX} = \frac{A}{B} + \frac{C \cdot 100}{A}$$

где

- А — количество слов в тексте,
- В — количество предложений в тексте,
- С — количество слов длинее 6 букв.



Task Type	1-60	61-70	71-80	81-90	91-100	101-110	111-120	121-130	131-140	141-150	151-160	161-170	171-180	181-190	191-200	201-210	211-220	221-230	231-240	241-250	251-260	261-270	271-280	281-290	291-300	301-310	311-320	321-330	331-340	341-350	351-360	361-370	371-380	381-390	391-400	401-410	411-420	421-430	431-440	441-450	451-460	461-470	471-480	481-490	491-500	501-510	511-520	521-530	531-540	541-550	551-560	561-570	571-580	581-590	591-600	601-610	611-620	621-630	631-640	641-650	651-660	661-670	671-680	681-690	691-700	701-710	711-720	721-730	731-740	741-750	751-760	761-770	771-780	781-790	791-800	801-810	811-820	821-830	831-840	841-850	851-860	861-870	871-880	881-890	891-900	901-910	911-920	921-930	931-940	941-950	951-960	961-970	971-980	981-990	991-1000																																																																																														
Test Type	1-60	61-70	61-70	71-80	71-80	81-90	81-90	91-100	91-100	101-110	101-110	111-120	111-120	121-130	121-130	131-140	131-140	141-150	141-150	151-160	151-160	161-170	161-170	171-180	171-180	181-190	181-190	191-200	191-200	201-210	201-210	211-220	211-220	221-230	221-230	231-240	231-240	241-250	241-250	251-260	251-260	261-270	261-270	271-280	271-280	281-290	281-290	291-300	291-300	301-310	301-310	311-320	311-320	321-330	321-330	331-340	331-340	341-350	341-350	351-360	351-360	361-370	361-370	371-380	371-380	381-390	381-390	391-400	391-400	401-410	401-410	411-420	411-420	421-430	421-430	431-440	431-440	441-450	441-450	451-460	451-460	461-470	461-470	471-480	471-480	481-490	481-490	491-500	491-500	501-510	501-510	511-520	511-520	521-530	521-530	531-540	531-540	541-550	541-550	551-560	551-560	561-570	561-570	571-580	571-580	581-590	581-590	591-600	591-600	601-610	601-610	611-620	611-620	621-630	621-630	631-640	631-640	641-650	641-650	651-660	651-660	661-670	661-670	671-680	671-680	681-690	681-690	691-700	691-700	701-710	701-710	711-720	711-720	721-730	721-730	731-740	731-740	741-750	741-750	751-760	751-760	761-770	761-770	771-780	771-780	781-790	781-790	791-800	791-800	801-810	801-810	811-820	811-820	821-830	821-830	831-840	831-840	841-850	841-850	851-860	851-860	861-870	861-870	871-880	871-880	881-890	881-890	891-900	891-900	901-910	901-910	911-920	911-920	921-930	921-930	931-940	931-940	941-950	941-950	951-960	951-960	961-970	961-970	971-980	971-980	981-990	981-990	991-1000	991-1000
Test Type	1-60	61-70	61-70	71-80	71-80	81-90	81-90	91-100	91-100	101-110	101-110	111-120	111-120	121-130	121-130	131-140	131-140	141-150	141-150	151-160	151-160	161-170	161-170	171-180	171-180	181-190	181-190	191-200	191-200	201-210	201-210	211-220	211-220	221-230	221-230	231-240	231-240	241-250	241-250	251-260	251-260	261-270	261-270	271-280	271-280	281-290	281-290	291-300	291-300	301-310	301-310	311-320	311-320	321-330	321-330	331-340	331-340	341-350	341-350	351-360	351-360	361-370	361-370	371-380	371-380	381-390	381-390	391-400	391-400	401-410	401-410	411-420	411-420	421-430	421-430	431-440	431-440	441-450	441-450	451-460	451-460	461-470	461-470	471-480	471-480	481-490	481-490	491-500	491-500	501-510	501-510	511-520	511-520	521-530	521-530	531-540	531-540	541-550	541-550	551-560	551-560	561-570	561-570	571-580	571-580	581-590	581-590	591-600	591-600	601-610	601-610	611-620	611-620	621-630	621-630	631-640	631-640	641-650	641-650	651-660	651-660	661-670	661-670	671-680	671-680	681-690	681-690	691-700	691-700	701-710	701-710	711-720	711-720	721-730	721-730	731-740	731-740	741-750	741-750	751-760	751-760	761-770	761-770	771-780	771-780	781-790	781-790	791-800	791-800	801-810	801-810	811-820	811-820	821-830	821-830	831-840	831-840	841-850	841-850	851-860	851-860	861-870	861-870	871-880	871-880	881-890	881-890	891-900	891-900	901-910	901-910	911-920	911-920	921-930	921-930	931-940	931-940	941-950	941-950	951-960	951-960	961-970	961-970	971-980	971-980	981-990	981-990	991-1000	991-1000
Test Type	1-60	61-70	61-70	71-80	71-80	81-90	81-90	91-100	91-100	101-110	101-110	111-120	111-120	121-130	121-130	131-140	131-140	141-150	141-150	151-160	151-160	161-170	161-170	171-180	171-180	181-190	181-190	191-200	191-200	201-210	201-210	211-220	211-220	221-230	221-230	231-240	231-240	241-250	241-250	251-260	251-260	261-270	261-270	271-280	271-280	281-290	281-290	291-300	291-300	301-310	301-310	311-320	311-320	321-330	321-330	331-340	331-340	341-350	341-350	351-360	351-360	361-370	361-370	371-380	371-380	381-390	381-390	391-400	391-400	401-410	401-410	411-420	411-420	421-430	421-430	431-440	431-440	441-450	441-450	451-460	451-460	461-470	461-470	471-480	471-480	481-490	481-490	491-500	491-500	501-510	501-510	511-520	511-520	521-530	521-530	531-540	531-540	541-550	541-550	551-560	551-560	561-570	561-570	571-580	571-580	581-590	581-590	591-600	591-600	601-610	601-610	611-620	611-620	621-630	621-630	631-640	631-640	641-650	641-650	651-660	651-660	661-670	661-670	671-680	671-680	681-690	681-690	691-700	691-700	701-710	701-710	711-720	711-720	721-730	721-730	731-740	731-740	741-750	741-750	751-760	751-760	761-770	761-770	771-780	771-780	781-790	781-790	791-800	791-800	801-810	801-810	811-820	811-820	821-830	821-830	831-840	831-840	841-850	841-850	851-860	851-860	861-870	861-870	871-880	871-880	881-890	881-890	891-900	891-900	901-910	901-910	911-920	911-920	921-930	922-930	931-940	931-940	941-950	941-950	951-960	951-960	961-970	961-970	971-980	971-980	981-990	981-990	991-1000	991-1000
Test Type	1-60	61-70	61-70	71-80	71-80	81-90	81-90	91-100	91-100	101-110	101-110	111-120	111-120	121-130	121-130	131-140	131-140	141-150	141-150	151-160	151-160	161-170	161-170	171-180	171-180	181-190	181-190	191-200	191-200	201-210	201-210	211-220	211-220	221-230	221-230	231-240	231-240	241-250	241-250	251-260	251-260	261-270	261-270	271-280	271-280	281-290	281-290	291-300	291-300	301-310	301-310	311-320	311-320	321-330	321-330	331-340	331-340	341-350	341-350	351-360	351-360	361-370	361-370	371-380	371-380	381-390	381-390	391-400	391-400	401-410	401-410	411-420	411-420	421-430	421-430	431-440	431-440	441-450	441-450	451-460	451-460	461-470	461-470	471-480	471-480	481-490	481-490	491-500	491-500	501-510	501-510	511-520	511-520	521-530	521-530	531-540	531-540	541-550	541-550	551-560	551-560	561-570	561-570	571-580	571-580	581-590	581-590	591-600	591-600	601-610	601-610	611-620	611-620	621-630	621-630	631-640	631-640	641-650	641-650	651-660	651-660	661-670	661-670	671-680	671-680	681-690	681-690	691-700	691-700	701-710	701-710	711-720	711-720	721-730	721-730	731-740	731-740	741-750	741-750	751-760	751-760	761-770	761-770	771-780	771-780	781-790	781-790	791-800	791-800	801-810	801-810	811-820	811-820	821-830	821-830	831-840	831-840	841-850	841-850	851-860	851-860	861-870	861-870	871-880	871-880	881-890	881-890	891-900	891-900	901-910	901-910	911-920	911-920	921-930	922-930	931-940	931-940	941-950	941-950	951-960	951-960	961-970	961-970	971-980	971-980	981-990	981-990	991-1000	991-1000
Test Type	1-60	61-70	61-70	71-80	71-80	81-90	81-90	91-100	91-100	101-110	101-110	111-120	111-120	121-130	121-130	131-140	131-140	141-150	141-150	151-160	151-160	161-170	161-170	171-180	171-180	181-190	181-190	191-200	191-200	201-210	201-210	211-220	211-220	221-230	221-230	231-240	231-240	241-250	241-250	251-260	251-260	261-270	261-270	271-280	271-280	281-290	281-290	291-300	291-300	301-310	301-310	311-320	311-320	321-330	321-330	331-340	331-340	341-350	341-350	351-360	351-360	361-370	361-370	371-380	371-380	381-390	381-390	391-400	391-400	401-410	401-410	411-420	411-420	421-430	421-430	431-440	431-440	441-450	441-450	451-460	451-460	461-470	461-470	471-480	471-480	481-490	481-490	491-500	491-500	501-510	501-510	511-520	511-520	521-530	521-530	531-540	531-540	541-550	541-550	551-560	551-560	561-570	561-570	571-580	571-580	581-590	581-590	591-600	591-600	601-610	601-610	611-620	611-620	621-630	622-630	631-640	631-640	641-650	641-650	651-660	651-660	661-670	661-670	671-680	671-680	681-690	681-690	691-700	691-700	701-710	701-710	711-720	711-720	721-730	721-730	731-740	731-740	741-750	741-750	751-760	751-760	761-770	761-770	771-780	771-780	781-790	781-790	791-800	791-800	801-810	801-810	811-820	811-820	821-830	821-830	831-840	831-840	841-850	841-850	851-860	851-860	861-870	861-870	871-880	871-880	881-890	881-890	891-900	891-900	901-910	901-910	911-920	911-920	921-930	922-930	931-940	931-940	941-950	941-950	951-960	951-960	961-970	961-97						

Результат:

readability

23 признаков

слова

5 признаков

“похожесть”

1 признак

Baseline



train 0.7
test 0.15
val 0.15

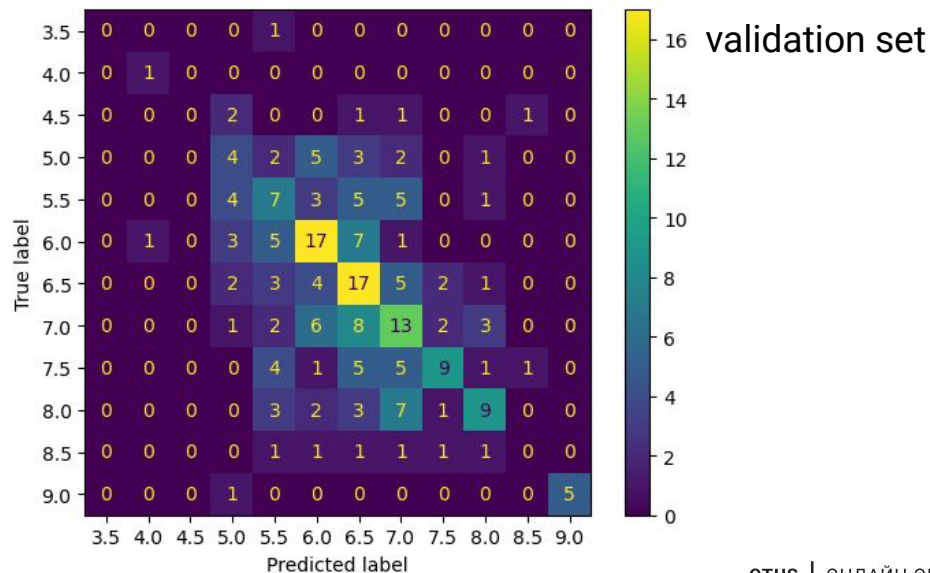
stratify=data['Overall']

	test	val
accuracy_score	0.395	0.381

данные: ['Essay', 'Question', 'Task_Type']

модель: LogisticRegression

	precision	recall	f1-score	support
3.5	0.00	0.00	0.00	1
4.0	0.50	1.00	0.67	1
4.5	0.00	0.00	0.00	5
5.0	0.24	0.24	0.24	17
5.5	0.25	0.28	0.26	25
6.0	0.44	0.50	0.47	34
6.5	0.34	0.50	0.40	34
7.0	0.33	0.37	0.35	35
7.5	0.60	0.35	0.44	26
8.0	0.53	0.36	0.43	25
8.5	0.00	0.00	0.00	6
9.0	1.00	0.83	0.91	6
accuracy			0.38	215
macro avg	0.35	0.37	0.35	215
weighted avg	0.39	0.38	0.37	215



Влияние новых признаков

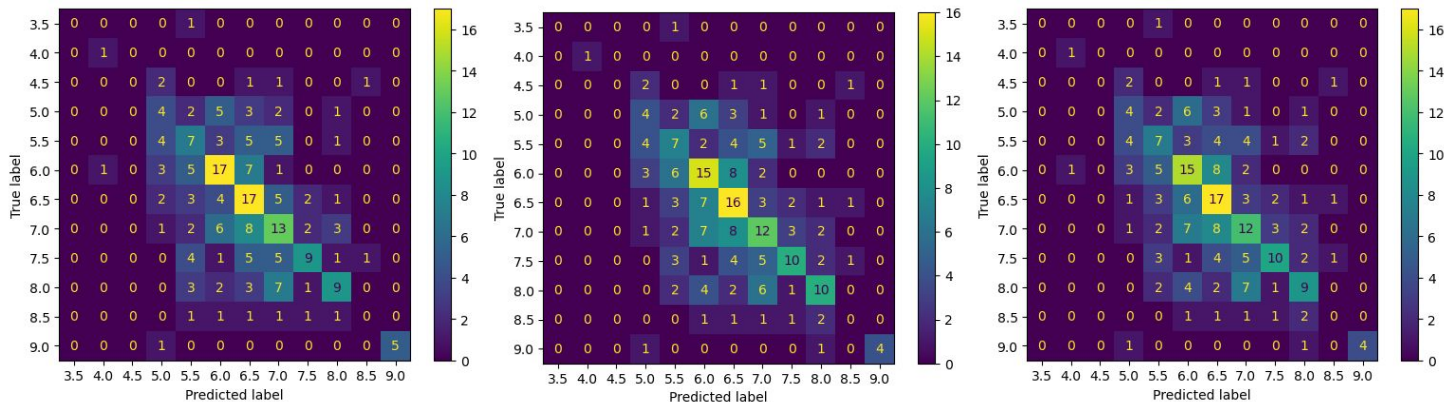


данные: все признаки vs. нескоррелированные признаки

модель: LogisticRegression

accuracy_score

	baseline	все данные	без скоррелированных данных
test	0.395	0.395	0.391
val	0.381	0.367	0.367



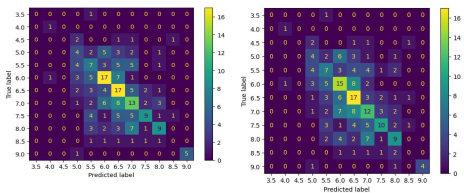
Влияние выбора модели



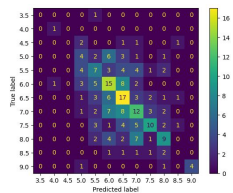
данные: нескоррелированные признаки

accuracy_score

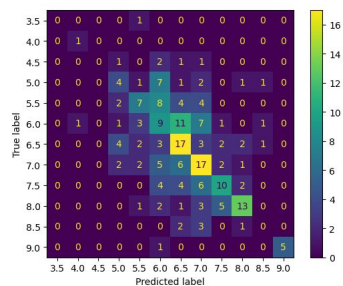
	baseline	LogisticRegression	Gradient Boosting	Random Forest	SVC	CatBoost	Voting Classifier
test	0.395	0.391	0.423	0.433	0.395	0.405	0.405
val	0.381	0.367	0.386	0.451	0.386	0.479	0.414
время обучения	5.6 s	4.23 s	2min 15s	2.71 s	2.25 s	36.7 s	2min 37s



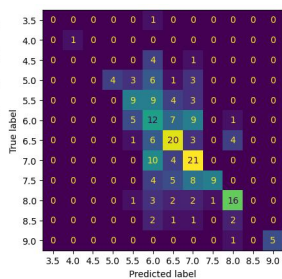
baseline



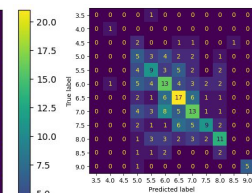
LogisticRegression



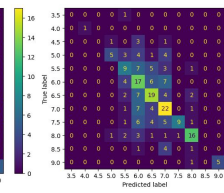
GradientBoosting



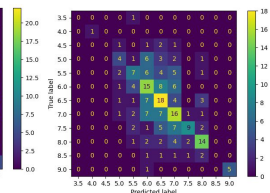
RandomForest



SVC



CatBoost



Voting Classifier



Влияние выбора данных



модель: VotingClassifier

accuracy_score	baseline	чистые признаки	все признаки	начальные признаки	простые признаки	сложные признаки
test	0.395	0.405	0.428	0.409	0.414	0.391
val	0.381	0.414	0.423	0.405	0.409	0.437

модель: RandomForestClassifier

accuracy_score	baseline	чистые признаки	все признаки	начальные признаки	простые признаки	сложные признаки
test	0.395	0.433	0.386	0.381	0.386	0.391
val	0.381	0.451	0.465	0.367	0.437	0.442



Проверка модели



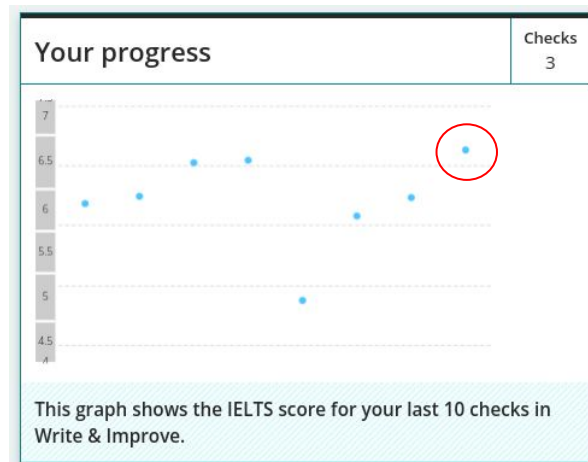
<https://writeandimprove.com>

Результаты модели RandomForestClassifier:

['6.0', '7.0']

Результаты модели VotingClassifier:

['7.0', '6.5']



Выводы



1. Даже сравнительно простые модели могут дать хорошие результаты при достаточном количестве данных
2. Корреляция в данных иногда не вредна
3. Ансамбли из нескольких моделей позволяют получать более стабильный результат

Спасибо за внимание!