Выпускная квалификационная работа

Разработка прогностической модели для оптимизации процесса набора абитуриентов, на основе анализа баллов ЕГЭ

Выполнил: Корякин Денис Дмитриевич

Руководитель ВКР: Семендяев Родион Юрьевич, к.т.н.

Актуальность работы

Ежегодное проведение приемных кампаний в ВУЗы

Многоэтапность процесса и участие одновременно в нескольких конкурсах создает для абитуриентов необходимость мониторинга большого количества данных для принятия верного решения.

Наличие издержек для абитуриентов и ВУЗов

Большая психоэмоциональная нагрузка и риск упущенных возможностей для абитуриентов.

Наличие дополнительных вступительных испытаний усложняет оценку вероятности успешного зачисления, в том числе за счет сложных доп. условий.

Риски недобора или дисбаланса по направления подготовки для ВУЗов.

Data-driven подход, оптимизирующий процесс выбора и зачисления

Основанные на данных инструменты, позволяющие повысить объективность принятие решения о предоставлении оригиналов документов.

Особенности зачисления в ведомственные ВУЗы

Иные принципы формирования конкурсных списков

«кандидаты, отнесенные по результатам профессионального психологического отбора к третьей категории профессиональной пригодности, располагаются в конкурсном списке после кандидатов, отнесенных к первой и второй категориям профессиональной пригодности, независимо от полученной суммы баллов.»

Приказ Министра обороны РФ от 7 апреля 2015 г. N 185
«Об утверждении Порядка и условий приема в образовательные организации высшего образования, находящиеся в ведении Министерства обороны Российской Федерации»

Обзор предметной области

Инструменты машинного обучения

- **Алгоритмы классификации и регрессии** — Алгоритмы логистической регрессии, случайного леса и XGBoost
- **Нейронные сети** — Используются в случаях наличия в датасетах большого количества параметров и сложных нелинейных параметров
- **Рекомендательные системы** — Базируются на алгоритмах коллаборативной фильтрации и могут создавать индивидуальные рекомендации
- **Методы анализа временных рядов** — Могут использоваться ВУЗами для планирования стратегии в среднесрочной перспективе на основе анализа исторических данных

Постановка цели и задач

Цель работы

Разработка прогностических моделей на основе анализа датасета с историческими данными которые:
прогнозируют вероятность зачисления абитуриента;
прогнозируют вероятность успешного поступления на имеющие специальности.

Задачи работы

01

Получение данных из БД

Формирование обезличенного датасета

02

Предварительная обработка данных

Устранение пропусков, балансировка и кодирование категориальных данных

03

Разведочный анализ данных

Оценка корреляции между признаками и их значимости

04

Бинарная модель классификации

Оценка вероятности зачисления

05

Многоклассовая модель классификации

Выбор оптимальной специальности

06

Тестирование и оценка модели

Тестирование разработанных моделей

Формирование датасета из БД

PostgreSQL

SQL-запрос	Подготовка к импорту в среду разработки	Преобразование в DateFrame
• Подготовка SQL-запроса для выгрузки необходимых полей за несколько лет • Исключение выгрузки персональной информации • Кодирование id и специальностей	**Подготовка Excel-файла** Инкапсуляция выгруженных данных в Excel-файл **Загрузка в среду разработки** Загрузка датасета в IDE	 Преобразование датасета в DateFrame для дальнейшей обработки

Предварительная оценка датасета

Получение описательных статистик

Оценка количества записей, выявление количества пустых записей, определение типов записей.

```
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Пол                 6165 non-null   object
 1   Статус              6165 non-null   object
 2   ПричинаОтчисления   3067 non-null   object
 3   Русский язык        5487 non-null   float64
 4   Математика          5649 non-null   float64
 5   Физика              5405 non-null   float64
 6   ФП                  5123 non-null   float64
 7   Профпригодность     5269 non-null   float64
 8   Специальность       6165 non-null   int64
dtypes: float64(5), int64(1), object(3)
memory usage: 481.6+ KB
```

Преобразование типов полей для последующей обработки

Преобразование категориальных данных в строковый тип.
Преобразование числовых данных в int.

	0
Пол	object
Статус	object
ПричинаОтчисления	object
Русский язык	int64
Математика	int64
Физика	int64
ФП	int64
Профпригодность	int64
Специальность	int64

dtype: object

Завершение обработки DataFrame

Обработка и замена NULL значений, удаление выбросов и ошибочных записей.

Визуализация распределения признаков

Данные категориальных признаков

Данные сбалансированы.

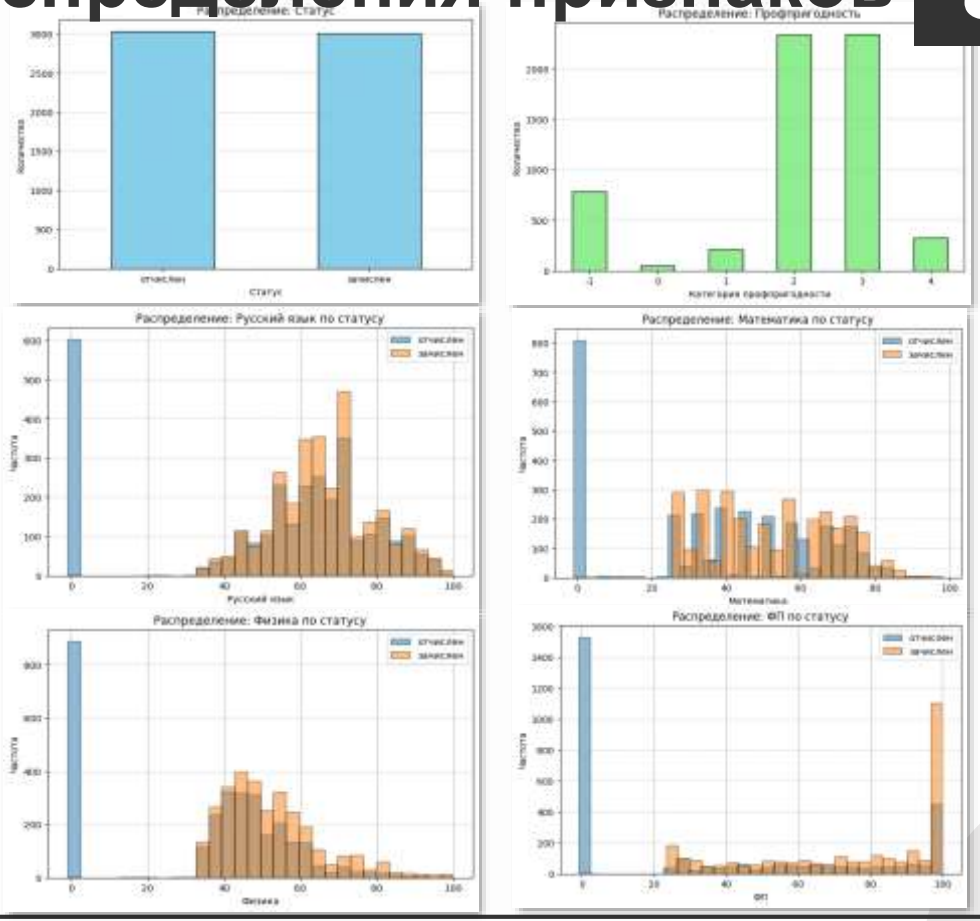
Кодирование категориальных признаков с помощью LabelEncoder

Данные числовых признаков

Распределены правдоподобно.

Минимальные и максимальные значения выбросов не имеют.

Результаты академических испытаний объединены в новый признак.

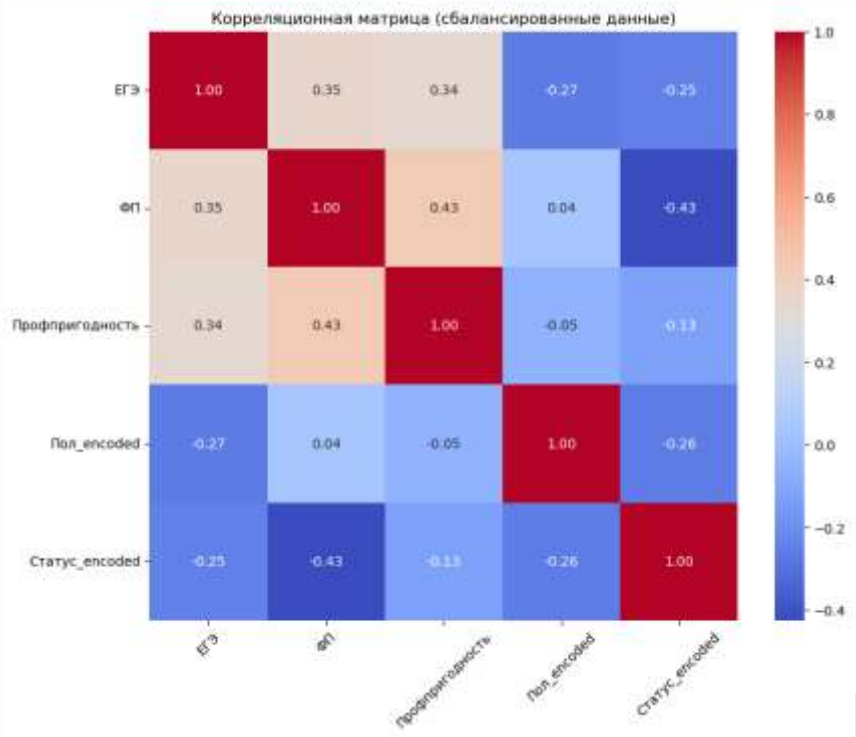


Построение корреляционной матрицы

Анализ матрицы

Проведена оценка величины корреляции на сбалансированных и кодированных данных.

Высокие уровни корреляции между факторными признаками отсутствуют, что позволяет предположить отсутствие проблемы мультиколлинеарности.



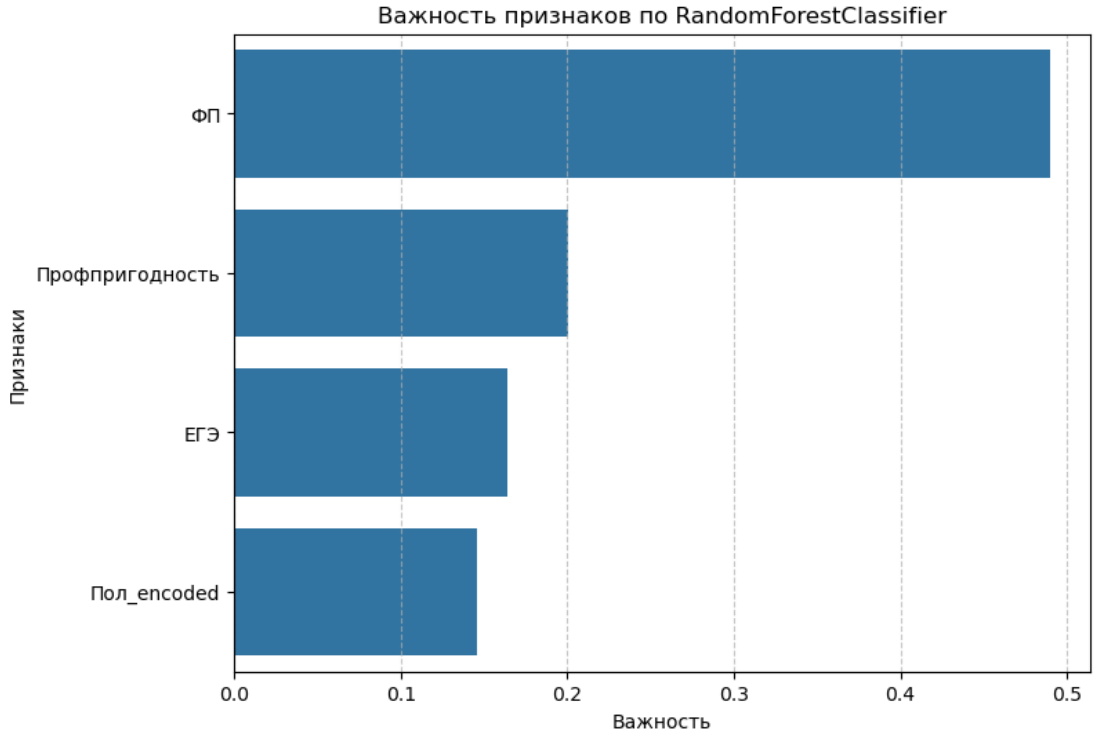
Оценка важности признаков

Оценка важности признаков

Оценка важности признаков проведена на основе алгоритма дерева решений.

Модель показала высокую точность: 99.8% на тестовых данных. Это указывает на то, что выбранные признаки хорошо разделяют целевые классы.

Оставляем все признаки для начальной версии модели, поскольку даже менее значимые признаки могут улучшить обобщение.

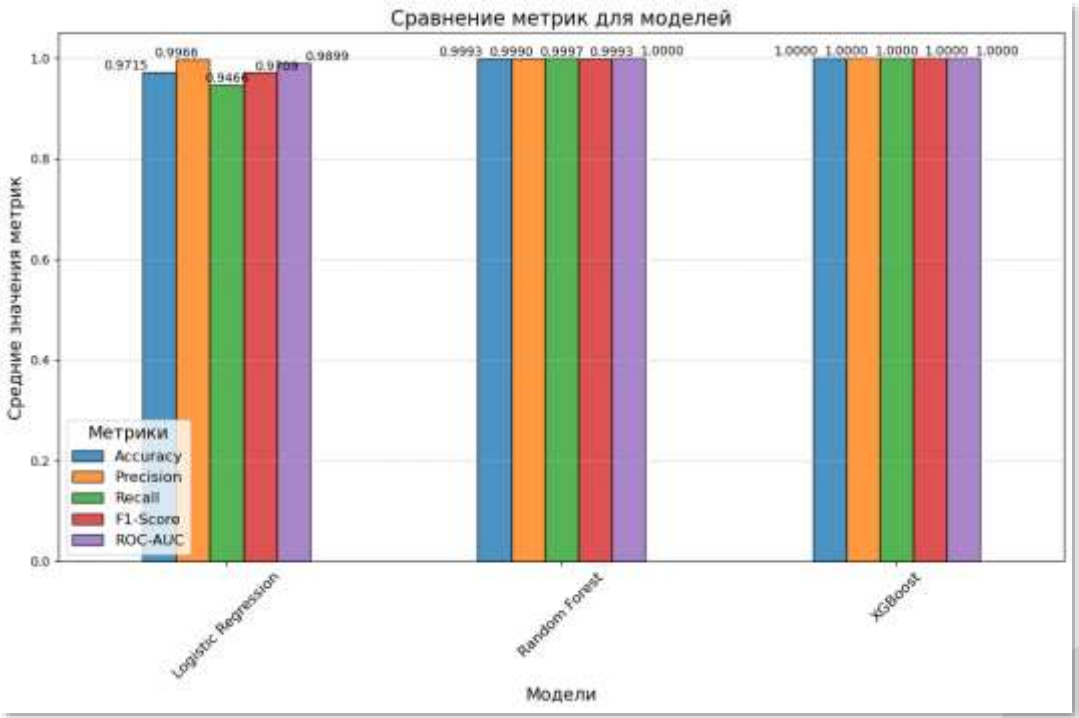


Построение модели бинарной классификации

Сравнение метрик

При сравнении метрик на кросс-валидации по построенным моделям наилучшие результаты показывает алгоритм XGBoost.

Тестирование на данных условного абитуриента показывает высокую правдоподобность результатов прогнозирования.

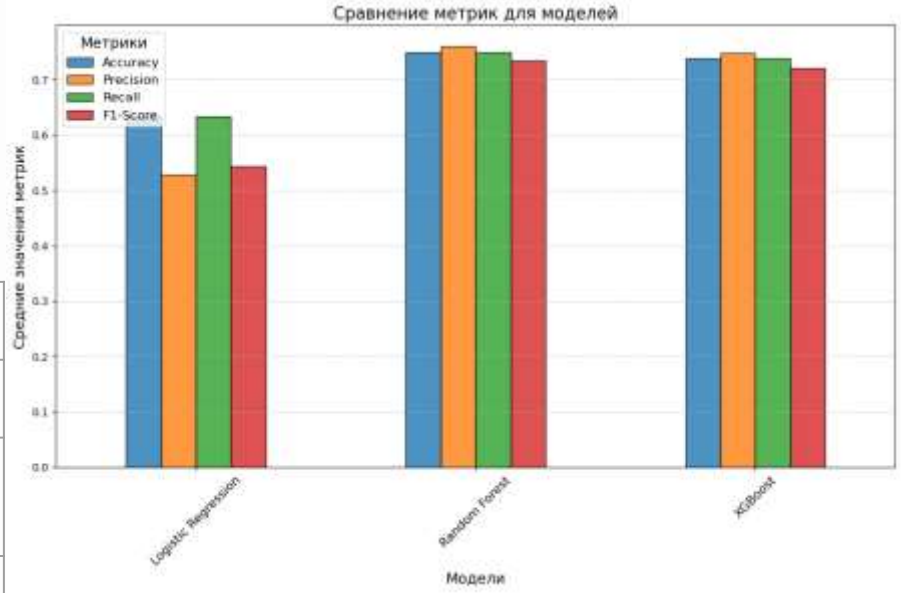


Построение прогностической модели для выбора специальностей

Поиск параметров по сетке

Для нахождения оптимальной модели выполнен поиск гиперпараметров по сетке

Модель	Лучшие параметры, подобранные по сетке	Accuracy	Precision	Recall	F1-Score
Логистическая регрессия	{'C': 10}	63.40%	52.92%	63.40 %	54.38%
Случайный лес	{'max_depth': 20, 'min_samples_leaf': 1, 'n_estimators': 200}	75.08%	76.12%	75.08 %	73.54%
XGBoost	{'learning_rate': 0.2, 'max_depth': 10, 'n_estimators': 200}	73.88%	74.52%	73.88 %	72.11%



Построение прогностической модели для выбора специальностей

Тестирование данных

Построена матрица ошибок, которая показывает удовлетворительную точность прогнозирования.

Тестирование моделей на данных условных абитуриентов

№ теста	ЕГЭ	Физическая подготовленность	Профпригодность	Пол	Вероятность зачисления	Расчет по специальностям
1	220	100	2	Муж	99,95%	32: 70.54% 33: 24.93% 31: 2.06% 23: 1.30% 22: 0.85% 11: 0.11% 24: 0.09% 21: 0.06% 12: 0.06%
2	228	89	2	Жен	64,31%	33: 100.00% 11: 0.00% 12: 0.00% 21: 0.00% 22: 0.00% 23: 0.00% 24: 0.00% 31: 0.00% 32: 0.00%
3	228	89	4	Муж	0,25%	0%

Заключение

Подготовлен и обработан датасет для создания прогностической модели

Сформирован, загружен и проанализирован датасет, содержащий данные об абитуриентах. Проведена обработка и балансировка данных, для последующего построения моделей.

Созданы прогностические модели процесса набора абитуриентов

Созданы, оптимизированы и протестированы модели бинарной и многоклассовой классификации, показавшие удовлетворительные результаты при прогнозировании вероятности поступления и подбора специальностей.

Выполнение целей и задач

Поставленные цели и задачи работы выполнены.

Практическая применимость достигнута

Ансамбль моделей может быть обучен на датасетах соответствующих ВУЗов, имеющих аналогичную модель поступления. Абитуриенты и ВУЗы получают дополнительный инструмент контроля за ходом приемной кампании.