ПРОГНОЗ УСПЕВАЕМОСТИ ПО ДАННЫМ ПЛАТФОРМ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

Владова А.Ю.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
Финансовый университет при Правительстве РФ, Москва, Россия
Avladova@ipu.ru

Аннотация. Предлагаемый подход к прогнозу успеваемости отличается тем, что прогнозную регрессионную модель обучают на преобразованных во временной ряд, нормализованных и зашумленных разновременные оценках активностей каждого учащегося, собранных платформой электронного обучения.

Ключевые слова: академическая успеваемость, статистический анализ, регрессия, траектории обучения, e-learning платформа.

Введение

Семестровая успеваемость студентов оценивается уровнем выполнения аудиторных, самостоятельных, контрольных и курсовых работ, а также активностью на семинарах. Выполнение и мониторинг оценок по этим работам автоматизируют с помощью платформ электронного обучения. Платформа электронного обучения — это набор сервисов, который позволяет пользователям получать доступ к образовательным курсам, учебным материалам и ресурсам [1]. Кроме прочего, платформа собирает данные о поведении пользователей, таких как процент завершения курса, время, затраченное на каждый модуль, уровень тестов и др. Кроме того, возможен сбор демографической информации (имя, адрес электронной почты и должность), местоположение и количество кликов, прокруток и поисковые запросы. Эти данные помогают администраторам платформ улучшить процесс обучения и предоставлять персонализированные рекомендации. Наряду с известными достоинствами таких платформ существует и ряд недостатков: ограниченная интерактивность; технические проблемы; ограниченная обратная связь; неэффективное мотивирование; ограниченный набор прививаемых навыков, а также недостаточность аналитики, предлагающей базовые инструменты и отчеты, а также неадекватные или недостаточные метрики. Слабым местом платформ электронного обучения является недостаточное использование методов прогнозирования для упреждающего выявления неуспевающих студентов.

Таким образом, сформулирована цель предлагаемой работы – по текущим активностям, зафиксированным платформой электронного обучения, спрогнозировать группу неуспевающих студентов.

1. Анализ литературы

Исследователи [2] взяли оценки по пяти курсам весеннего семестра и для генерации прогнозов использовали четыре алгоритма регрессионного машинного обучения: случайный лес (RF), байес (BR), адаптивный бустинг (AdaBoost) и экстремальный градиентный бустинг (XGBoost). Результаты показали, что ансамблевая модель для RF и XGBoost показала наилучшие результаты. Среди всех оцененных активностей электронного обучения, тесты оказали значительное влияние на прогнозирование успеваемости студентов. Автор [3] на основе оценок аспирантов при обучении в аспирантуре предсказывает вероятность защиты кандидатской диссертации, используя деревья решений. В статье [5] прогноз целевой переменной выполнен для каждого временного ряда с помощью модели Хольта экспоненциального сглаживания, реализованных в библиотеке Statsmodels, учитывающей тенденцию изменения данных. В статье [7] на основе значительного количества независимых от эксперта разновременных оценок генерируют динамические признаки, характеризующие изменение оценок каждого учащегося в течении семестра по видам активностей и за счет них добиваются высокой точности прогноза.

2. Структура данных

Последовательность оценок после переименования признаков и обезличивания студентов имеет вид временного ряда (табл. 1).

Таблица 1. Фрагмент исходных данных

Индекс студента	**1**	**2**	**3**	**…**	**15**	**16**	**17**
ГАf5	0	0	7	…	25	38	0
ВКf5	0	0	10	…	60	80	0
ВДm5	2	0	13	…	55	92	0

Источник: преобразованные автором данные платформы электронного обучения Финансового университета при Правительстве РФ

Из таблицы следует, что оценки выставлены в разных шкалах. По нормированным оценкам построены распределения (рис. 1)

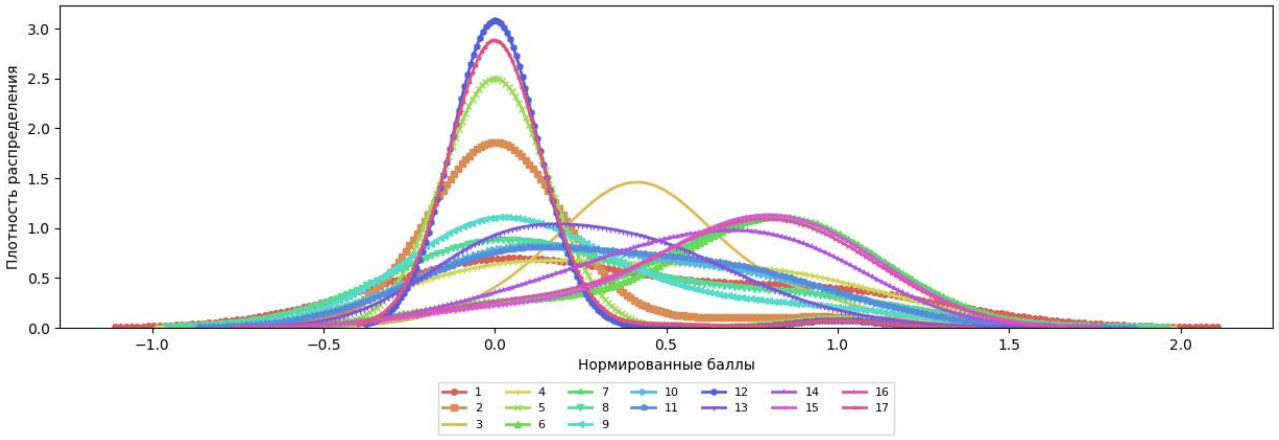

Рис. 1. Распределения нормированных данных

Такие признаки, как 1, 2, 3, 6, 12 имеют распределение близкое к нормальному. Матожидание части распределений признаков, например 14, 15, 16 смещено вправо относительно 0.

3. Математическая постановка задачи

Задача прогнозирования итоговой оценки учащегося может быть сформулирована как задача регрессии, где мы стремимся предсказать числовое значение итоговой оценки на основе значений различных признаков успеваемости учащегося. Пусть n - количество учащихся, m - количество признаков успеваемости каждого учащегося, $x_i = (x_{i1}, x_{i2}, ..., x_{im})$ - вектор признаков успеваемости для i-го учащегося, где x_{ij} представляет значение j-го признака для i-го учащегося. Требуется построить модель $f(x_i)$, которая сможет предсказывать итоговую оценку y_i для i-го учащегося на основе значений его признаков успеваемости, то есть $y_i = f(x_i) + e_i$, где y_i - фактическое значение итоговой оценки, $f(x_i)$ - предсказанное значение, и e_i - ошибка модели.

Чтобы построить модель $f(x_i)$ можно использовать различные методы регрессии, такие как линейная регрессия, метод ближайших соседей, регрессионные деревья, нейронные сети [4, 5].

4. Прогноз итоговой успеваемости учащегося

Для предсказания итоговой оценки учащегося на основе значений признаков успеваемости выполнены пять шагов.

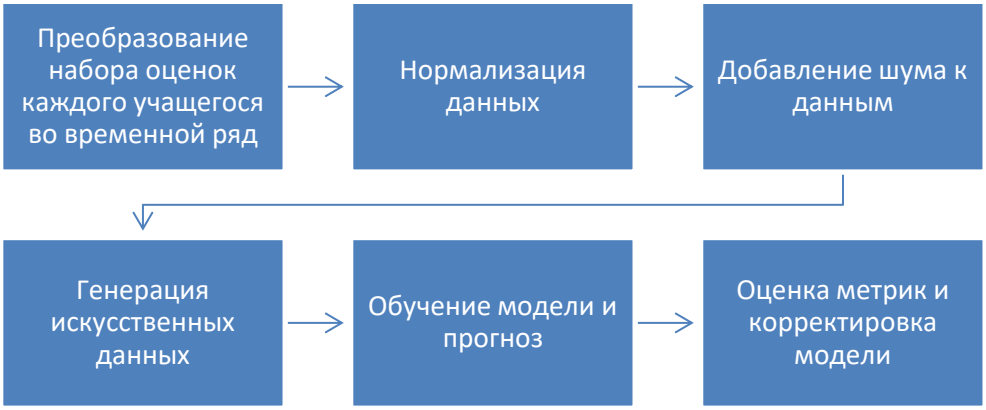

Рис. 2. Этапы прогнозирования итоговой оценки

На первом шаге выполнить преобразование набора оценок каждого учащегося во временной ряд, в котором каждому признаку датасета, содержащему оценку за определенную активность, дается хронологический номер выставления оценки. Это позволяет учитывать динамику успеваемости студента. На втором шаге провести нормализацию данных для обеспечения сопоставимости значений различных признаков. На третьем шаге добавить к данным шум, распределённый по стандартному нормальному закону. Добавление небольшого уровня шума к данным помогает модели учиться на более реалистичных данных, что может улучшить ее обобщающую способность. На четвертом шаге сгенерировать искусственные данные на основе уже имеющихся для предотвращения переобучения модели прогнозирования. На пятом шаге обучить модель и выполнить прогноз итоговой оценки каждого учащегося. В заключении оценить метрики, такие как Mean Squared Error (MSE), Mean Absolute Error (MAE), коэффициент детерминации (R squared) и скорректировать параметры модели.

4.1. Подготовка данных

Преобразование набора оценок каждого учащегося во временной ряд проведено примемением словаря feature_mapping к датафрейму quantitative_data, включающему только количественные признаки [6] исходного датафрейма df (рис. 3)

```
quantitative_data = df.select_dtypes(include=['int64', 'float64'])
feature_mapping = {
    'Активность 1': 1,
    'Программирование 1': 2,
    'Дом.задание 1': 3,
……
    'Экзамен': 15,
    'Итог': 16,
    'Пересдача': 17}
whiskers_data =
df_merged.rename(columns=feature_mapping).set_index('Индекс').sort_index(axi
s=1)
```

Рис. 3. Преобразование во временной ряд

Такое преобразование не учитывает неравномерность отметок времени проведения активностей и может влиять на точность прогноза. Нормализация данных датафрейма whiskers_data в интервал [0, 1] проведена с помощью `MinMaxScaler`. Нормализованные значения признаков преобразованы в новый датафрейм whiskers_normalized, с исходными названиями столбцов и индексами (рис. 4).

```
scaler = MinMaxScaler(feature_range = (0, 1))
whiskers_normalized_values = scaler.fit_transform(whiskers_data)
whiskers_normalized = pd.DataFrame(whiskers_normalized_values,
columns=whiskers_data.columns, index=whiskers_data.index)
```

Рис. 4. Нормализация данных

Полученный набор данных разделен в пропорции 80/20 на обучающий и тестовый. Причем в обучающий и в тестовый наборы попали студенты, не сдавшие экзамен. Результаты прогнозирования итоговых оценок (красная точка) по набору нормализованных оценок каждого студента (синие точки) линейной регрессией выглядят следующим образом (рис. 5).

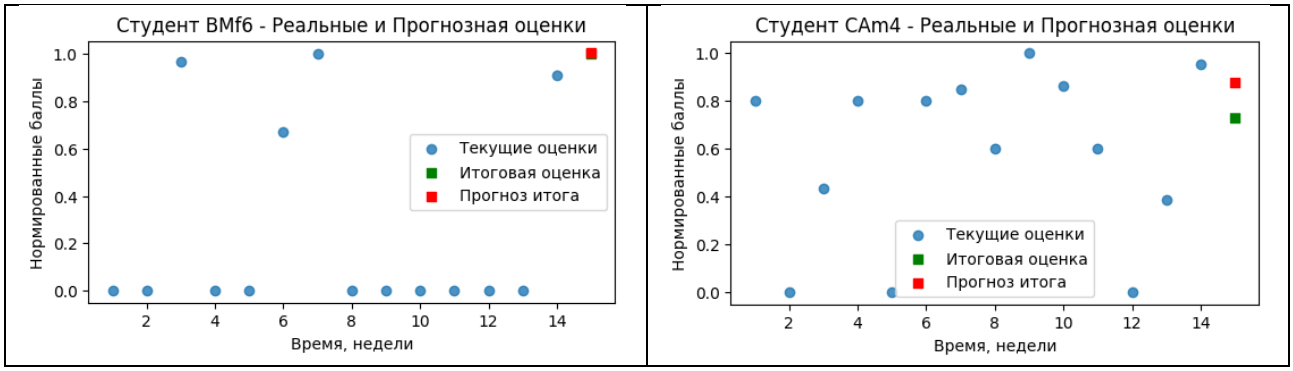

Рис. 5. Сравнение реальных и прогнозных оценок студента: а) совпадение, б) несовпадение

Метрики прогнозной модели на нормализованных данных составляют: Mean Absolute Error: 0.0725, Mean Squared Error: 0.0093, Root Mean Squared Error: 0.096, R squared: 0.9087. Для избегания переобучения модели к данным добавлен нормальный шум с помощью кода рис. 6.

```
from sklearn.preprocessing import FunctionTransformer
def add_noise(X):
    noise = np.random.normal(0, 0.1, X.shape)  # Generate random noise
    return X + noise
noisy_transformer = FunctionTransformer(add_noise)
X_train_noisy = noisy_transformer.transform(X_train)
X_test_noisy = noisy_transformer.transform(X_test)
linear_regression_model(X_train_noisy, X_test_noisy, y_train, y_test)
```

Рис. 6. Зашумливание данных

Метрики прогнозной модели на зашумленных данных составляют: Mean Absolute Error: 0.1126, Mean Squared Error: 0.0178, Root Mean Squared Error: 0.1335, R squared: 0.8255.

Эти результаты отражают потенциальную эффективность прогнозной модели и устойчивость к незначительному изменению в данных.

5. Заключение

Обзор литературы посвящен прогнозированию траекторий успеваемости учащихся с использованием методов машинного обучения. Недостатками практической реализации предлагаемых методов являются значительное количество качественных признаков при небольшой размерности пространства признаков и сильное влияние экспертов на оценки.

Средняя абсолютная ошибка прогнозов на зашумленных данных составляет 11,26 %, что означает, что модель в среднем ошибается на эту величину при прогнозировании. Среднее значение квадрата разности между фактическими значениями и прогнозами модели составляет всего 1,78 %. Значение коэффициента детерминации R^2 составляет 0.8255 и является мерой того, насколько хорошо прогнозы модели соответствуют фактическим значениям. Фактически, модель объясняет 82.55% дисперсии зависимой переменной.

Примеры данных и кода представлены на странице https://github.com/avladova/Student-performance-prediction/blob/main/MLSD24_Student_performance_Vladova%22.ipynb

Литература

1. *Qiu, F., Zhang, G., Sheng, X. et al.* Predicting students' performance in e-learning using learning process and behaviour data // Sci Rep. – 2022. – Vol. 12, - P. 453. https://doi.org/10.1038/s41598-021-03867-8.

2. *Malak A., Al-Ayyoub M., Shatnawi F., Rawashdeh S. Abbott R.* Predicting students' academic performance using e-learning logs // IAES International Journal of Artificial Intelligence. – 2023. - Vol. 12. – P. 831-839. doi: 10.11591/ijai.v12.i2.
3. *Клементьев А.А.* Использование прогнозных методов оценки успеваемости аспирантов на основе данных LMS-платформ // Общество: социология, психология, педагогика. - 2020. - № 12.
4. *Владова А.Ю.* Формирование пространства признаков и авторегрессионных моделей для прогноза отступлений железнодорожного полотна // Проблемы управления. - 2023. - № 2. - С. 54-64. doi: 10.25728/pu.2023.2.5.
5. *Владова А.Ю., Владов Ю.Р.* Прогноз температуры грунта трассы линейного протяженного объекта // Безопасность труда в промышленности. - 2020. - № 6. - С. 14-20. doi: 10.24000/0409-2961-2020-6-14-20.
6. Цифровизация математики в вузе / Л. Р. Борисова, В. А. Бывшев, А. Ю. Владова [и др.]. – Москва: ООО "Издательство Прометей". - 2021. – 578 с.
7. *Владова А.Ю.* Формирование групповой и индивидуальной траекторий успеваемости по данным e-learning платформ // Управление большими системами: Выпуск 111. М.: ИПУ РАН, 2024. - С. 179-196. doi: 10.25728/ubs.2024.111.7.