Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка алгоритмов экстракции признаков данных для задач прогнозирования

Выполнила: Прокопенко Надежда, гр. 7383
Руководитель: Геппенер В.В., д.т.н., профессор
Консультант: Шевская Н.В.

Санкт-Петербург, 2021

Цель и задачи

Актуальность: востребованность в маркетинговой сфере, где прогнозируется результат продвижения услуги. **Проблема:** сложность в выявлении признаков, влияющих на принятие решения, высокое время обработки большого количества признаков.

Цель: разработать алгоритм экстракции признаков для прогнозирования данных, имеющие количество признаков больше 10.

Задачи:

- Изучить существующие методы экстракции признаков данных;
- сформировать требования к разрабатываемым алгоритмам;
- разработать алгоритмы снижения размерности для дальнейшего прогнозирования данных;
- реализовать консольное приложение для демонстрации результатов работы алгоритмов;
- оценить метрики разработанных алгоритмов.

Существующие методы экстракции признаков данных

Таблица 1 – Сравнение аналогов

Методы	Сложность расчетов	Чувствительность к выбросам	Min кол-во компонентов	Точность, %
ICA	Низкая	+	2	64.02
PCA	Низкая	+	2	76.18
LDA	Высокая	-	1	86.31

Требования к разрабатываемым алгоритмам

- Снижение размерности данных до минимально-необходимой;
- удобство визуализации;
- точность прогнозирования выше 85%.

Разработка алгоритмов (1)

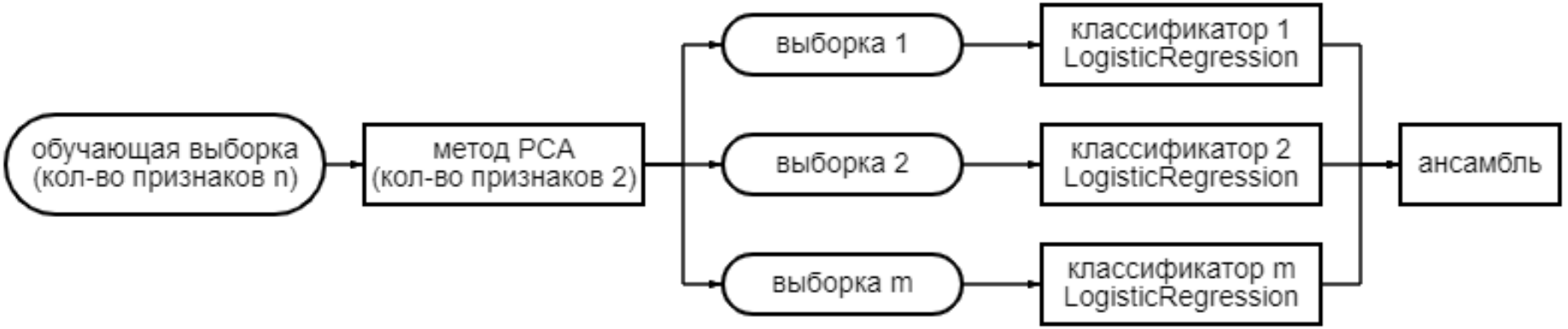

Рисунок 1 – Схема работы алгоритма прогнозирования ансамблем с использованием снижения размерности пространства признаков PCA, где m задается пользователем

Разработка алгоритмов (2)

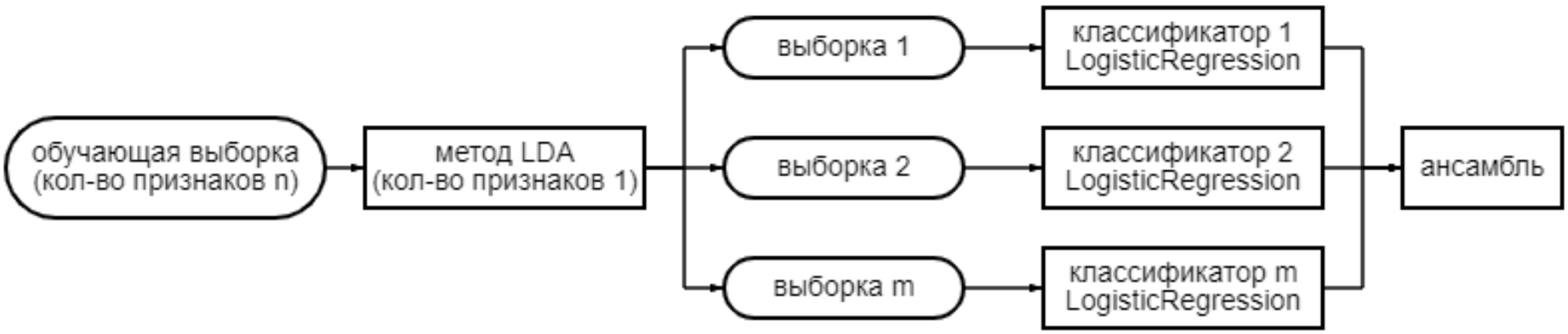

Рисунок 2 – Схема работы алгоритма прогнозирования ансамблем с использованием снижения размерности пространства признаков LDA, где m задается пользователем

Приложение для демонстрации работы алгоритмов

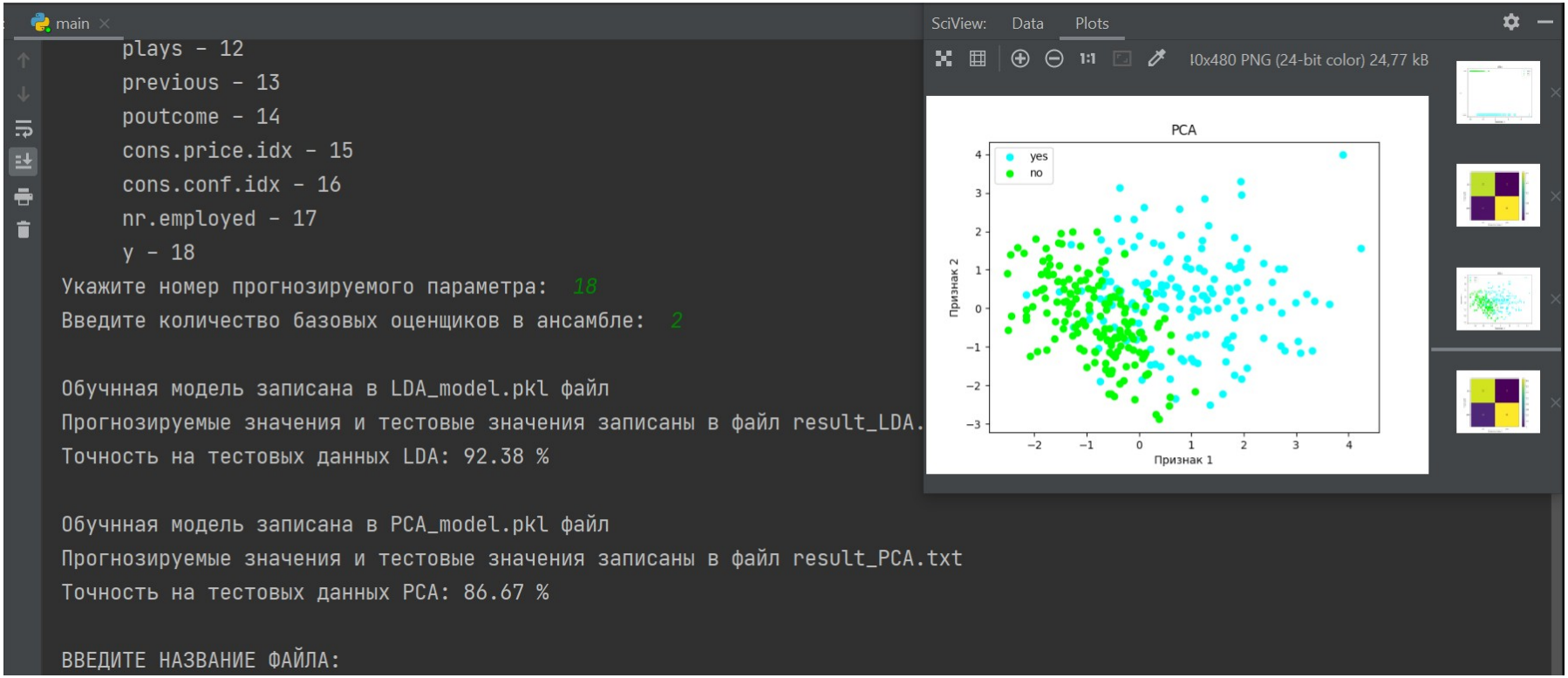

Рисунок 3 – Пример запуска консольного приложения для демонстрации результатов работы программы

Метрики разработанных алгоритмов

Таблица 2 – Сравнение результатов работы алгоритмов

Метод снижения размерности	Размерность пространства признаков	Метод прогнозирования	Точность прогнозирования
PCA	2	LogisticRegression	76.18%
		BaggingClassifier (ансамбль)	86.67%
LDA	1	LogisticRegression	86.31%
		BaggingClassifier (ансамбль)	92.38%

Визуализация данных после применения алгоритмов

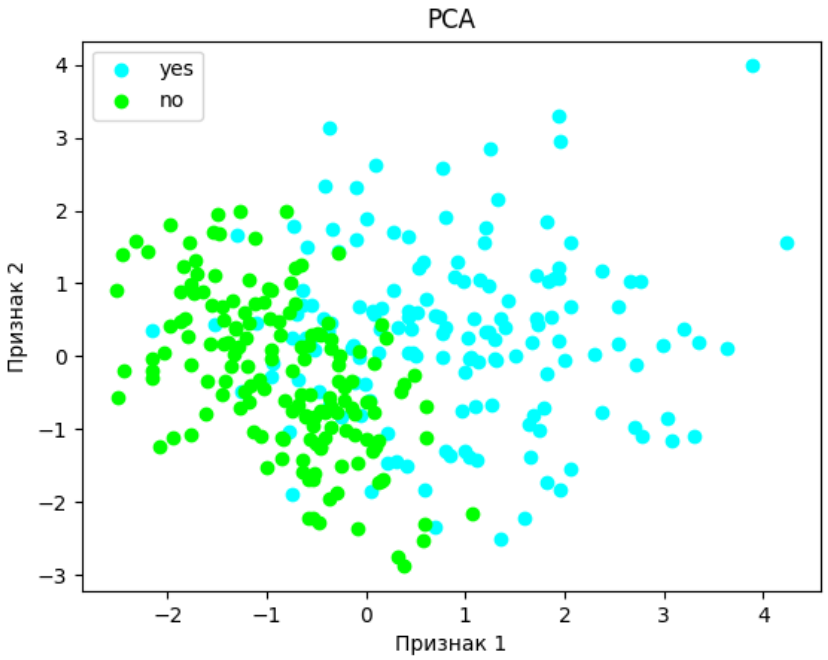

Рисунок 4 – Визуализация данных в двумерном пространстве после применения метода экстракции признаков PCA

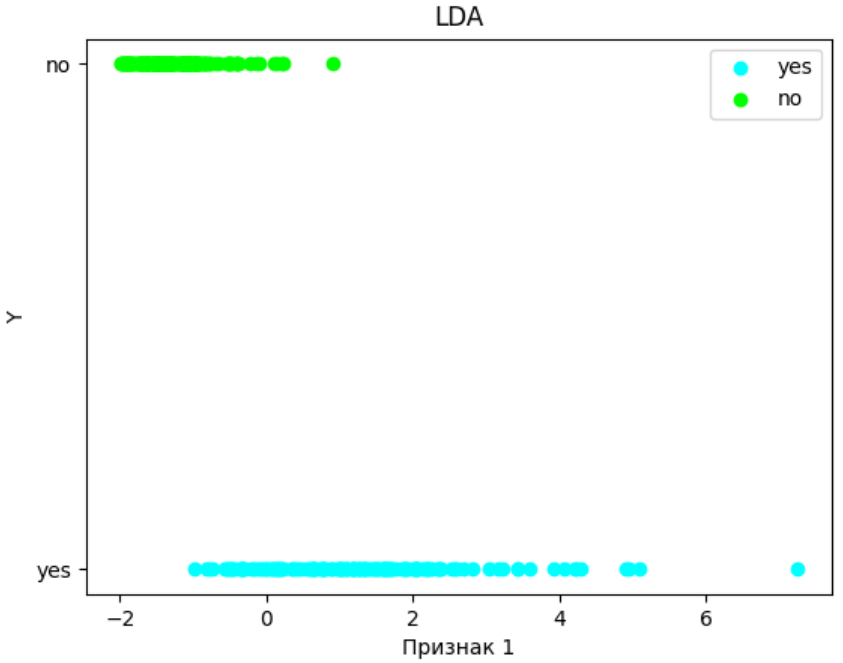

Рисунок 5 – Визуализация данных после применения метода экстракции признаков LDA

Заключение

- При сравнении аналогов выявлены самые точные методы экстракции признаков для задач прогнозирования (PCA, LDA)
- Сформулированы требования к разрабатываемым алгоритмам: снижение размерности данных до минимально-необходимой, удобство визуализации, точность выше 85%.
- Разработаны алгоритмы с точностью прогнозирования 86.67%(PCA) и 92.38%(LDA) с использованием техники создания ансабля.
- Создано консольное приложение, позволяющее пользователю настраивать количество базовых оценщиков в ансабле и прогнозировать результаты для данных с числом признаков более 10.
- Оценены метрики разработанных алгоритмов, полученные метрики соответствуют требованиям.

Дальнейшие направления исследований включают в себя повышение точности прогнозирования данных с использованием экстракции признаков для задач регрессии

Апробация работы

- Репозиторий проекта https://github.com/ProkopenkoNadezhda/diploma.git.

Запасные слайды

Основные определения

В машинном обучении снижение размерности — это преобразование данных, состоящее в уменьшении числа переменных путём получения главных переменных.

Выделение признаков (экстракция признаков) — это разновидность абстрагирования, процесс снижения размерности, в котором исходный набор исходных переменных сокращается до более управляемых групп (признаков) для дальнейшей обработки, оставаясь при этом достаточным набором для точного и полного описания исходного набора данных.

Размерность — количество независимых параметров, необходимых для описания состояния объекта, или количество степеней свободы системы.

Ансамбль — алгоритм, который состоит из нескольких алгоритмов машинного обучения.