


**Федеральное государственное автономное образовательное учреждение  
высшего образования  
«Национальный исследовательский университет «Высшая школа экономики»  
Факультет компьютерных наук  
Образовательная программа «Прикладная математика и информатика»  
Направление подготовки 01.03.02 «Прикладная математика и информатика»  
бакалавриат**

## **О Т Ч Е Т по преддипломной практике**

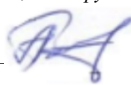
**Выполнил студент гр. БПМИ 194**

**Соколов Ян Олегович**  
(ФИО)

  
(подпись)

**Проверил:**

**Романенко Алексей Александрович**  
(должность, ФИО руководителя практики)

  
(подпись)

**23.04.2023**

(дата)

**2023 год**

## **Содержание:**

- 1. Цели и задачи практики**
- 2. Постановка задачи**
- 3. Актуальность темы**
- 4. Основные сложности задачи**
- 5. Обзор существующих методов решения поставленной задачи**
- 6. Выбор методов решение. Обосновании выбора.**
- 7. План решение поставленной задачи**
- 8. Полученные результаты**
- 9. Выводы**
- 10. Обзор литературы**

### **Цели и задачи практики**

Моя задача состояла в том, чтобы подготовить различные пайплайны по сегментации обучающей выборки продакт-бейзд на базе кластеризации, продакт-бейзд на основе бизнес-иерархии, а также выстроить фреймворк обучения алгоритмов прогнозирования спроса в рамках подготовленных пайплайнов, сравнить точность алгоритмов для различных пайплайнов сегментации обучающей выборки.

### **Постановка задачи**

Главной задачей проекта является исследование эффективности сегментации обучающей выборки при построении прогнозирующего пайплайна с помощью ML моделей. В рамках этой задачи было решено рассмотреть различные типы разбиения тренировочной выборки и сравнения их между собой.

### **Актуальность темы**

Область прогнозирования спроса становится все более востребованной в последние годы в связи с необходимостью точного прогнозирования потребительского спроса в различных отраслях. На данный момент не существует универсального решения в таких задачах, каждая из задач рассматривается с точки зрения уже существующих методов, выбирая лучший. Мы хотим представить новый подход в решении подобных задач - обобщенная сегментация обучающей выборки. Исследуя этот метод, мы стремимся улучшить такие аспекты как точность и надежность моделей прогнозирования спроса, которые имеет решающее значение для принятия решений в таких отраслях, как розничная торговля, логистика и управление поставками.

### **Основные сложности задачи**

Для корректного сравнения двух пайплайнов: с использованием сегментации и без требуется исследовать большой объем данных, >12млн. строк, поэтому одной из основных проблем является количество и сложность вычислительных операций.

Также стоит отметить, что при определенных разбиениях обучающей выборки, в некоторых случаях модели обучаются на относительно небольшом объеме данных, что может повлечь за собой недообучение моделей и снижение их предсказательной способности.

### **Обзор существующих методов решения поставленной задачи.**

Проанализировав литературу связанную с нашей темой, было выявлено, что авторы статей [1, 2, 3] не провели достаточно глубокий анализ данной сферы, они используют примитивные методы разбиения обучающей выборки для обучения нескольких моделей, а также не приводят информацию об объеме данных, на которых они тестируют свои пайплайны. Мы же собираемся исследовать данную область, путем построения различных сложных пайплайнов сегментирования обучающей выборки на реальных данных большого объема.

## **Выбор методов решения. Обоснование выбора.**

На данный момент для исследования данной задачи было решено построить различные пайплайны по сегментированию обучающей выборки: product-based clustering - кластеризация по принципу схожести товаров в разных магазинах, store-based clustering - кластеризация магазинов по сходству спроса в них, city-based clustering - разделение обучающей выборки с учетом расположения данного магазина.

Также для исследования различных пайплайнов сегментации было решено проверить подход построения индивидуальных моделей для каждой пары товар-магазин и сравнить с пайплайном, состоящим из одной модели.

## **План решения поставленной задачи**

Входные данные для нашей задачи - датасет состоящий из более 12 млн строк еженедельных данных о спросе в конкретном магазине по конкретному товару. Основные колонки в нашем датасете - product\_rk, store\_location\_rk, location, demand, week.

Данные разбиваем на обучающую и тестовую выборку - на обучающей получаем сегментацию и обучаем модели, на тестовой выборке строим предсказания и таким образом сравниваем наши пайплайны.

Для сегментации нашей обучающей выборки мы решили использовать алгоритм кластеризации для иерархического типа данных - AgglomerativeClustering. С его помощью мы можем тестировать наши различные подходы по разбиению обучающей выборки. Качество разбиения на кластеры и подбор оптимального их количества мы вычисляем путем одновременного учета двух индексов - Silhouette score и Davide Bouldine score.

После получения меток кластеров по обучающей выборке добавляем их в тестовую выборку и обучаем для каждого кластера свою модель, затем получаем предсказания на тестовой выборке и сравниваем пайплайны.

Для пайплайна по построению индивидуальной модели для каждой пары товар-магазин был выбран порог минимального количества семплов, требуемого для обучения каждой модели. Для тех пар, количество семплов которых было меньше заданного порога мы решили обучить отдельную модель.

В качестве моделей мы используем модель градиентного бустинга CatBoostRegressor. Для каждой из них мы подбираем гиперпараметры с помощью написанной нами функции. Признаки, которые используются в обучении этих моделей - набор лаговых характеристик, которые мы рассчитали с помощью написанной нами функции, позволяющей создать разные признаки с разными свойствами путем перебора параметров (размер окна, лага, агрегирующей функции, колонок-фильтров).

## **Полученные результаты.**

На данный момент мы построили несколько пайплайнов: кластеризация по среднему спросу по разным уровням месторасположения (регионам, городам, ...), магазинам и товарам. Мы обучили отдельные модели для каждого из разбиений и одну общую модель на всей обучающей выборке. Ни один из использованных методов не дал

улучшений в качестве предсказаний, ошибки на тестовой выборке отличаются менее чем на 1%.

## **Выводы.**

За время прохождения практики я рассмотрел пайплайны с различными методами сегментации обучающей выборки и сравнил их с пайплайном содержащим лишь одну модель обученную на всей выборке.

Дальнейшим шагом является улучшение текущих методов сегментации обучающей выборки и поиск хитрого разбиения пар товар-магазин на кластеры, с помощью которых получится добиться уменьшения ошибки на тестовой выборке в сравнении с единственной обученной моделью.

## **Список изученной литературы.**

[1] S. Mazzanti, "What Is Better: One General Model or Many Specialized Models?"

[2] C. Short, "Segmentation: Does it Impact Predictive Modelling Results for People Analytics?"

[3] Y. Hadar, "Should I Train a Model for Each Customer or Use One Model for All of My Customers?"