

**ЗАДАНИЕ ДЕМОНСТРАЦИОННОГО ЭКЗАМЕНА**

|  |  |
| --- | --- |
| **Код комплекта оценочной документации** | КОД 1.2-2023-2025 |
| **Год действия задания** | 2023 |
| **Номер варианта задания** | 3 |

**ОПИСАНИЕ ЗАДАНИЯ**

**Описание модуля 1: «****Парсинг и предобработка данных»**

На демонстрационном экзамене вам предстоит разработать модель машинного обучения, предлагающую лучшую следующую покупку клиенту банка, для увеличения доли покупок с данной карты в общей структуре трат клиента и повышения лояльности.

Основное направление при продаже товаров и услуг­ – это история про встраивание в контекст клиента. Решение должно представлять собой поток контекстно-обусловленных выгодных предложений (рекомендаций) клиенту по следующей покупке.

Решение должно учитывать потребности всех заинтересованных сторон. Банку это решение должно повысить количество транзакций, который совершает клиент банка. А клиент получает возможность экономить, а также и расширять свой потребительский опыт получая предложения, основанные на опыте похожих людей. Возможности по гео-таргетированным предложениям (вплоть до real-time) дают дополнительную контекстную силу подходу.

Решение может представлять рекомендацию льготных покупок карточным клиентам банка с указанием места совершения покупки, срока действия предложения, учитывающую потребности клиента и приоритеты банка.

Предложение должно удовлетворять следующим основным критериям:

* предложение должно быть интересно и релевантно клиенту (если он ходит в самые бюджетные продуктовые магазины – то получает предложения на самую бюджетную аптеку, а не дорогую; если пользуется онлайн магазинами еды – то предложение на онлайн магазин одежды и пр.)
* предложение обязательно должно расширять интенсивность использования карты клиентом (увеличивать среднее число и объем транзакций в месяц). То есть клиент через предложение должен научиться либо платить в новых товарных категориях, либо платить чаще, либо тратить больше.

В рамках всего задания вам потребуется предобработать данные, выполнить анализ данных и выявить ключевые зависимости, построить необходимые модели машинного обучения, разработать приложение и API для пользователей сервиса.

**При выполнении модуля 1 ставятся следующие цели:**

1. Подготовка набора данных для дальнейшего построения прогнозирующей модели;

**При выполнении данного модуля 1 ставятся следующие задачи:**

1. Выполнить парсинг данных для сбора информации о покупках клиентов банка;
2. Формирование структуры набора данных;
3. Провести предварительную обработку данных;
4. Выполнить построение и отбор признаков;
5. Выполнить кластеризацию данных

**Требования к оформлению письменных материалов**

Письменный материал отсутствует.

**Представление результатов работы**

Результат выполнения Модуля 1«Парсинг и предобработка данных»: предобработанные данные (архив Data.zip), отчет о проделанной работе (Report\_C1.html, Report\_C1.ipynb), дополнительные комментарии коду (Readme.txt).

**Необходимые приложения**

Необходимые приложения смотреть в папке «КОД 1.2 Приложения к вариантам».

Приложение 1: Исходные данные (Data.zip)

**ЗАДАНИЕ**

|  |
| --- |
| 1.1 Парсинг данных |
| На основании файлов Data1.csv и Data2.csv необходимо построить исходный набор данных. Набор данных должен быть преобразован в единый файл формата .csv.  Следует дополнить набор какими-либо другими данными, если они могут быть полезны для дальнейшего исследования. |

|  |
| --- |
| 1.2 Формирование структуры набора данных |
| Задача определения рекомендации покупок заключается в определении класса (кластера) покупателей на основе информации о покупателях и истории их покупок.  Решение должно учитывать потребности банка в увеличении продаж и интересы покупателей. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на определение таких классов (кластеров), и оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора документов. |

|  |
| --- |
| 1.3 Предварительная обработка данных |
| Предобработанный набор данных не должен содержать пропусков и аномалий (выбросов). Приведите данные к приемлемому формату. |

|  |
| --- |
| 1.4 Формирование дополнительных атрибутов |
| [В](https://gogov.ru/articles/covid-rt) представленных данных необходимо выделить геоданные, которые могут быть полезны для дальнейшего обучения моделей. В набор данных следует добавить координаты места (или возможных мест) совершенной покупки и любую другую информацию (регион, город), позволяющую определить место покупки. |

|  |
| --- |
| 1.5 Кластеризация |
| Выполнить кластеризацию данных по сходству покупателей несколькими способами (не менее трех). Выберите метрику оценки качества кластеризации. Обоснуйте выбор методов и приемов. Выполните визуальный анализ кластерных структур и оценки качества кластеризации. Определите лучший алгоритм кластеризации на основе выбранной метрики. |

|  |
| --- |
| 1.6 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip. |

**Описание модуля 2:** **«Построение, обучение и оптимизация модели»**

В этом модуле продолжается работа с данными, подготовленными в предыдущей сессии. Требуется осуществить выбор алгоритма классификации, построить модель и провести оптимизацию полученной модели машинного обучения в контексте исследуемой задачи.

**При выполнении модуля 2 ставятся следующие цели:**

Построение классификатора предпочтений пользователей.

**При выполнении модуля 2 ставятся следующие задачи:**

1. Провести разведочный (визуальный) анализ данных;
2. Выполнить разбиение выборки на обучающую и валидационную;
3. Осуществить построение моделей классификации;
4. Оценить качество полученных моделей в соответствии со спецификой решаемой задачи;
5. Выполнить оптимизацию лучшей модели.

**Требования к оформлению письменных материалов**

Письменный материал отсутствует.

**Представление результатов работы**

Результат выполнения Модуля 2 «Построение, обучение и оптимизация модели»: предобработанные данные (архив Data.zip), отчет о проделанной работе (Report\_C1.html, Report\_C1.ipynb), дополнительные комментарии коду (Readme.txt).

**Необходимые приложения**

Необходимые приложения смотреть в папке «КОД 1.2 Приложения к вариантам».

Приложение 1: Исходные данные (Data.zip)

**ЗАДАНИЕ**

|  |
| --- |
| 2.1 Разведочный анализ |
| Проведите анализ плотности распределения атрибутов и целевой переменной набора данных. Дайте интерпретацию полученных результатов.  Используя программные средства, визуализируйте зависимости атрибутов в наборе данных. Визуализация должна отражать влияние атрибутов на категории, суммы покупок и предпочтения пользователей по покупкам. Произведите расчеты зависимостей по выбранным алгоритмам. Приведите интерпретацию полученных результатов. |

|  |
| --- |
| 2.2 Разбиение выборки |
| Выполните разбиение полученной выборки на обучающую и тестовую. Проведите обучение моделей, основанных на различных алгоритмах. Сделайте предсказание категории товара или услуги на тестовой выборке. Выполните оценку моделей разной степени сложности в соответствии с выбранной метрикой. Определите модель, показавшую лучшее качество. |

|  |
| --- |
| 2.3 Оптимизация модели |
| Выполните настройку полученной модели уменьшив вычислительную сложность модели (выбор значимых признаков, понижение размерности и т.д.). Путём преобразования набора данных, добейтесь более точной работы выбранной модели. Опишите приемы генерации новых данных и результаты, к которым они привели, рассматривая все ранее определенные показатели качества. Оцените качество полученной модели, сделайте вывод.  Необходимо оптимизировать полученную модель под решаемую задачу, настраивая гиперпараметры. Выполните оценку данной модели после настройки гиперпараметров.  Построить для данной модели кривые валидации и обучения, интерпретируйте полученные результаты. |

|  |
| --- |
| 2.4 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip. |