**Министерство образования и науки Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**“САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ,**

**МЕХАНИКИ И ОПТИКИ”**

Список и описание шаблонов документов, которые могут использоваться для описания результатов соответствующих этапов ВКР

Выполнила: Павленко А.Д

Группа: К3443

Санкт-Петербург  
2020

**Документы, которые будут использоваться в дальнейшей работе:**

1. Business understanding (CRISP-DM) (Понимание бизнеса)
2. Data understanding (CRISP-DM) (Понимание данных)
3. Modeling (CRISP-DM) (Построение модели)
4. **Business understanding (CRISP-DM) (Понимание бизнеса)**

В зависимости от специфики конкретного проекта, какие-то разделы могут быть опущены или изменены.

* 1. **Контекст проекта**

В этом разделе представлен общий обзор контекста проекта. Здесь указывается, в какой предметной области осуществляется проект, какие проблемы были выявлены и почему их можно эффективно решить с помощью интеллектуального анализа данных.

**1.2. Терминология**

Раздел «Терминология» позволяет людям, которые не знакомы с предметной областью и проблемами проекта, изучить понятия, используемые в работе над проектом, и лучше понять контекст проекта.

(1) Словарь бизнес-терминов, которые формируют доступное для проекта понимание бизнеса.

(2) Словарь терминов анализа данных, иллюстрированные примерами, относящимися к бизнес-задаче, о которой идет речь.

**1.3. Бизнес-цели и критерии их достижения**

В разделе описываются цели проекта с точки зрения бизнеса. Для каждой цели должны быть предусмотрены критерии её достижения, то есть чёткие метрики для определения того, были ли достигнуты цели в результате выполнения проекта. В этом разделе также должны быть перечислены цели, которые были рассмотрены, но отклонены. Кроме того, должно быть представлено обоснование бизнес-целей.

Таблица 1 – Бизнес-цели проекта

|  |  |  |  |
| --- | --- | --- | --- |
| Бизнес-цели | Задачи для достижения цели | Метрика определения успешности достижения цели | Временные рамки |
|  |  |  |  |
|  |  |  |  |

**1.4. Требования, предположения и ограничения**

В этом разделе перечисляются общие требования к выполнению проекта: тип результатов проекта, предположения относительно характера проблемы и используемых данных, а также ограничения, налагаемые на проект.

* Требования к проекту, включая график выполнения, ясность и качество результатов, безопасность, а также правовые вопросы.
* Предположения, сделанные в рамках проекта. Это могут быть предположения о данных, которые можно проверить в процессе анализа данных, но также могут включать непроверяемые идеи о бизнесе, на котором основывается проект.
* Ограничения, налагаемые на проект. Это могут быть ограничения на наличие ресурсов, но также могут включать технологические ограничения, такие как размер данных, которые целесообразно использовать для моделирования.

**1.5. Обзор аналогов**

Аналоги - это примеры компаний той же отрасли, в которой проводится следующий анализ, или другой отрасли, которая при этом имеет ключевое сходство. Может описываться как успешное достижение бизнес-цели, так и наоборот.

**1.6. Первоначальная оценка инструментов и методов**

Этот раздел дает начальное представление о том, какие инструменты и методы могут быть использованы и как. Он описывает требования к инструментам и методикам, перечисляет доступные инструменты и методики и сопоставляет их с требованиями.

1. **Data understanding (CRISP-DM) (Понимание данных)**

Цель шага – понять слабые и сильные стороны предоставленных данных, определить их достаточность, предложить идеи, как их использовать, и лучше понять процессы заказчика. Для этого мы строим графики, делаем выборки и рассчитываем статистики.

* 1. **Первоначальный отчет о сборе данных**

На данном этапе описывается, как различные источники данных были получены и извлечены.

Данных раздел может включать в себя:

* источники информации
* местонахождение данных и методы, которые использовали для их получения
* проблемы при получении данных
* решение проблем, возникших при получении данных
  1. **Описание данных**

В этом разделе должна быть описана полученная информация.

Необходимо описать количество данных (например, количество полей в каждой таблице), формат данных, идентификацию каждого из полей и другие функции. Так же проанализировать, соответствует ли полученная информация сформированным выше требованиям.

* 1. **Исследование данных**

Исследования данных включает в себя решение вопросов для интеллектуального анализа с использованием методов отчетности и визуализации данных.

Данный раздел может включать:

* выбор инструментов для интеллектуального анализа
* анализ ключевых статистик
* визуализация результатов
* предположения и выводы

Отчет включает в себя первоначальную гипотезу или первые выводы, влиянии на остальную часть проекта.

1. **Modeling (CRISP-DM) (Построение модели)**

Следующий шаг — обучение моделей. Как правило, оно выполняется итерационно – мы пробуем различные модели, сравниваем их качество, делаем перебор параметров и выбираем лучшую комбинацию.

* 1. **Предположения моделирования.**

В этом разделе определяются предположения, сделанные относительно данных, и любые предположения, которые подразумеваются в используемой технике моделирования.

* 1. **Выбор и описание модели.**

Необходимо определиться, какие модели будут использоваться. Выбор модели зависит от решаемой задачи, типов атрибутов и требований по сложности. При выборе следует ответить на следующие вопросы:

* Достаточно ли данных, поскольку сложные модели как правило требуют большей выборки?
* Сможет ли модель обработать пропуски данных (какие-то реализации алгоритмов умеют работать с пропусками, какие-то нет)?
* Сможет ли модель работать с имеющимися типами данных или необходима конвертация?

Далее подробно изучить и кратко изложить выбранную модель.

* 1. **Обучение модели.**

Запуск цикла обучения и после каждой итерации фиксируется результат.   
Далее, следует ответить и описать ответы на следующие вопросы:

* Показывает ли модель какие-то интересные закономерности?  
  Например, что точность предсказания на 99% объясняется всего одним атрибутом.
* Какова скорость обучения/применения модели?  
  Если модель обучается 2 дня, возможно, стоит поискать более эффективный алгоритм или уменьшить обучающую выборку.
* Были ли проблемы с качеством данных?   
  1. **Оценка результатов.**

В этом разделе описываются результаты тестирования модели/моделей, исходя из условий описания выше.

Задачи шага:

* провести технический анализ качества модели (ROC, Gain, Lift и т.д.),
* оценить, готова ли модель к внедрению в КХД (или куда нужно),
* достигаются ли заданные критерии качества,
* оценить результаты с точки зрения достижения бизнес-целей