**Структура презентации:**

1. **Презентация спикера (1 минута)**

• Кто я сейчас

• Чему я научился / Что я умею благодаря курсам

• Мои цели после обучения / Чему я обучался

**2 слайда / 100 слов / 168 слов**

*Здравствуйте, уважаемые члены комиссии и однокурсники!*

*Сегодня я рад представить вам итоговый проект, над которым работал по окончании всего курса.*

*Этот проект - результат многих бессонных ночей, проведенных за компьютером, и множества чашек кофе. Однако, я уверен, что все трудности были оправданы, ведь перед собой вы видите уникальное творение, сочетающее в себе знания, полученные на курсе, и мою собственную креативность.*

*Приступим!*

*Для начала немного о себе…*

*Меня зовут Наиль.*

*Я работаю специалистом по снабжению в региональном распределительном центре крупной российской компании.*

*Всегда питал интерес к программированию, но обычное написание кода казалось мне скучным и однообразным.*

*Именно data science показала мне, что кодирование тоже может быть интересным и творческим занятием.*

*Во время обучения на курсах я получил много полезных навыков:*

* *Улучшил свое аналитическое мышление,*
* *разобрался в прикладной математике,*
* *научился работать с разными источниками данных,*
* *решать задачи с помощью алгоритмов машинного обучения и программировать на Python.*

*Моя цель - погрузиться в мир машинного обучения и реализовать свои идеи без привязки к определенному офису или городу, что, несомненно, улучшит мою жизнь.*

1. **Цели и задачи проекта (2 минуты)**

• С каким проектом я работал

• Какой цели мне надо было добиться

• Каких изменений надо было достичь

**4 слайда / 180 слов / 181 слово**

*Цель данного проекта создать одну из моделей для оценки кредитного риска, задачей которой является предсказание выхода клиента в дефолт по кредиту, на основе имеющихся бинаризированных данных о клиенте.*

*Передо мной стояла задача - подготовить автоматизированный пайплайн с подобранной моделью, который по вызову fit будет готовить данные и обучать модель, а по вызову predict — делать предсказания на заданном наборе данных с качеством не менее 0,75 по метрике ROC-AUC.*

*1 – если клиент с большей вероятностью уйдет в дефолт по выданному кредиту.*

*0 – если дефолт по кредиту маловероятен*

*К реализации задачи были установлены четкие критерии оценки, которые вы видите на слайде.*

*Для обучения модели были предоставлены данные в виде паркета из 11 файлов и csv-файл с целевой переменной (вероятностью дефолта).*

*Данные содержат бинаризированную информацию о клиентах с кредитами, их истории и финансах, всего 61 атрибут.*

*Каждая запись в датасете представляет один конкретный кредитный продукт, выданный конкретному заёмщику. Одному заемщику может соответствовать несколько продуктов.*

*Файл с целевой переменной содержал только по одному значению для каждого заемщика. Что первоначально поставило в тупик, так как количество записей в данных превышало количество значений целевой переменной в несколько раз.*

1. **Реализация проекта (4 минуты) — пошаговое описание реализации.**

• Зачем я выполнял конкретный этап.

• Что я сделал.

• Как я это сделал.

• Результат выполнения этапа.

**8 сладов / 350 слов**

***Загрузка данных***

*Чтобы решить эту проблему, я решил закодировать все входные атрибуты с помощью метода get\_dummies из библиотеки pandas. В результате число столбцов в датасете возросло с 61 до 397.*

*Затем, путем группировки по ID клиента и суммирования полученных столбцов, я вытянул данные в векторы признаков для каждого ID клиента.*

*В итоге, каждому ID соответствует только одна строка в дата фрейме без потери данных, а количество строк равно числу значений целевой переменной.*

***Очистка данных***

*В процессе преобразования переменных в фиктивные, не было возможности использовать параметр drop\_first, чтобы удалить первые уровни. В результате в обучающую выборку попали зависимые переменные.*

*Чтобы решить эту проблему я решил удалить их с помощью цикла по списку непреобразованных названий атрибутов.*

*В процессе работы, цикл перебирает названия столбцов преобразованного датафрейма, в поиске столбцов, начинающихся с названия атрибута, как только находит первый из них – удаляет его и переходит к следующему атрибуту.*

***Выбор модели***

*При выборе модели основными критериями были скорость и качество обучения.*

*По итогу были отобраны следующие модели:*

*SVC*

*Random Forest Classifier*

*Extra Trees Classifier*

*Hist Gradient Boosting Classifier*

*В ходе экспериментов выбор был сделан в пользу последней. Эта модель показала наилучшие значения по скорости и качеству, за счет использования градиентного бустинга и улучшенного алгоритма для работы с большим количеством выборок.*

*В то же время линейная модель SVC не смогла дать результат, так как скорость обучения была слишком высокой, пришлось прервать эксперимент.*

1. **Результаты проекта (3 минуты)**

• Какие изменения произошли/должны произойти в результате применения предложений.

• На что это повлияет и как (желательно в цифрах).

**6 слайдов / 260 слов**