**Слайд 1.**

Я представляю работу на тему «Разработка и использование моделей классификации в кредитном скоринге (на примере ОАО «Белинвестбанк»)».

**Слайд 2.**

Сначала обсудим задачу, которая стала объектом данной работы;

Затем перейдем к идентификационной форме модели, использованной в этой работе;

Потом представим использованный набор данных и комплекс преобразований над ним для подготовки к моделированию;

Завершают мое выступления тонкости связанные с получением модели и её валидацией.

Соответствующие пункты вы можете найти в бумажных презентациях представленных вам. Маркер указывает на принадлежность к той или иной структурной части работы.

**Слайд 3.**

В Банк ежедневно поступает множество заявок **(анимация)** на выдачу кредита. Требуется проводить данных предоставленных клиентом для принятия решения о выдаче или удержании кредита. Когда речь касается физических лиц, а, на сегодняшний день, уже и малого бизнеса – объемы поступающих заявок настолько велики, что рассмотреть в ручную каждую заявку не представляется возможным. Для обработки поступающих заявок создается кредитный конвейер **(анимация)** – механизм который позволяет обрабатывать заявки автоматически. Сердцем которого становиться скоринговая модель **(анимация)** – механизм который позволяет автоматически принимать решение **(анимация)** о выдаче или удержании кредита для каждого потенциального кредитополучателя на основе наблюденной ранее статистики. Аналогичная схема представлена на рисунке 1.

Наша задача заключалась в формировании такой модели. Понятно, что поставленная задача сводиться к задаче классификации. И формально **(анимация)** может быть описана так: сформировать некоторое правило R которое на основании входных данных X выдает бинарный оклик кодирующий выдачу или удержание кредита.

**Слайд 4.**

Изначально для решения такой задачи, на данных, которые мы представим далее, мы использовали классический алгоритм логистической регресии. Тут, сверху представлена идентификационная форма такой модели. Для нашей модели качества такого классификатора было недостаточно. Заметим, что это получается линейный классификатор, покажем это на примере **(анимация)** тут слева два класса точек расположились по двум переменным, если на таких данных сроить логистическую регрессию получатся предсказания верностей как отложены на втором рисунке. Выбрав некоторую вероятность после которой мы считаем клиента склонным к дефолту у нас получается отложить линию уровня **(анимация)**, которая и станет дискриминирующей в данном случае. Как видно, даже в такой простой задаче модель допускает ошибки.

Наша идея заключалась в том, чтобы усложнить функцию под логитом, чтобы «сломать» эту прямую в нужном месте **(анимация)** – перейти к кусочно линейному классификатору и, за счет этого, добиться лучших предикативных свойств модели. Сравнение результата логистической регрессии и модели предложенной в работе представлено в таблице 1 в бумажной версии презентации.

**Слайд 5**

Для построения модели с описанными свойствами мы использовали идеи лежащие в нейронных сетях. Формально идентификационную форму такой модели удобно представлять в виде графа, как на рисунке 1, где дуга для каждого рассмотренного наблюдения передает некоторое число, а каждый узел представляет собой некоторое преобразование данных.

В каждом из преобразований присутствует сумматорная функция, которая представляет собой линейную комбинацию активаций предыдущего слоя – на графе это входящие в каждый нейрон дуги.

Нейронные сети принято условно разделять на слои на этом рисунке это никак не связанные между собой нейроны расположившиеся в одном столбце. Как отдельный слой **(анимация)** рассматривают входы модели – каждый из таких нейронов по исходящим из него дугам пускает значения соответствующих ему переменных. Слои никак не связанные с внешней средой называют скрытыми слоями, в нашем случае это ReLU преобразования. Слой который формирует отклики модели называется выходным слоем **(анимация)**, в общем случае, это не обязательно один нейрон но в нашем случае именно так. А заложенное преобразование уже знакомая нам логит функция.

**Слайд 6**

Перейдем к рассмотрению использованного набора данных и преобразований над ним. Был использован набор данных о течении кредитных отношений с физическими лицами, содержащий 247 062 наблюдения и 45 показателей. Все показатели а также их краткая характеристика представлены в таблице 2 предоставленного вам печатного материала. Там можно найти название каждого показателя, тип данных, область допустимых значений и число возможных уровней в случае номинативного показателя, последним столбцом будет число пропущенных значений в каждом столбце – это большая проблема рассматриваемого набора данных, во многих столбцах доля пропусков достигает 95%.

Особое внимание следует уделить столбцу “Число дней просрочки платежа” **(анимация)** далее мы будем использовать обозначение *Y*. В соответвии с политикой банка займ можно считать вышедшим в дефолт, если просрочка достигла срока 60 **(анимация)** и более дней таким образом был осуществлён переход к задаче бинарной классификации.

**Также следует уделить внимание тому факту, что в выборке присутствует всего n наблюдений с дефолтом что составляет n%. Критический недостаток наблюдений с дефолтом еще одна большая проблема с которой приходится сталкиваться в процессе моделирования.**

К обработке переменных разных типов несколько отличаются подходы потому **(анимация)** сразу разделим переменные на 3 типа: даты, числовые и номинативные. Для обработки дат сразу было применено следующее преобразование – для каждой даты было отчитано число дней от некоторой базовой даты, таким образом **(анимация)** они превратились в числовые переменные.

Далее мы исключили из рассмотрения 3 категориальные **(анимация)** переменные «Вид деятельности по ОКЭД», «Кредитный продукт» и «Место работы». Дело в том, что они нередко принимали очень много исключительных уровней. Конечно, они могли подлежать укрупнению, но для того надо было использовать дополнительные источники информации такие вещи выходят за границы данной работы.

Таблица содержала некоторые показатели которые не могут быть известны в момент оформления кредитополучателем контракта «отношение планового срока погашения задолженности к реальному», «причина прекращения договора» и «дата фактического закрытия договора».

Показателей осталось достаточно много, и мы решили попробовать их комбинировать для вычисления новых показателей. Например, отняв дату планируемого закрытия кредита от даты оформления кредитного договора можно было получить срок на который выдавался займ, а далее разделив на него сумму договора можно получить величину платежа в единицу времени и т.д. таким образом мы **(анимация)** добавили пять показателей два номинативных и три числовых.

**Слайд 7**

Далее было решено произвести очистку от невозможных значений и выборосов. Когда речь идет о невозможных значениях это какие-то ошибки заполнения или обработки. Например, по смыслу очевидно, что показатель «работа последнее место стаж лет» не может принимать отрицательных значений, однако в полученной таблице, как ни странно, такие наблюдения присутствуют. Такие и некоторые подобные случаи были попросту удалены из рассмотрения.

Выбросом называют результат измерения выбивающийся из общей выборки. Изучая описательные статистики числовых переменных, мы заметили ряд переменных статистики которых представлены на слайде или в таблице 3 раздаточного материала. Особое внимание **(анимация)** обращено на последние две строки. 75-я персентиль и максимальное значение. Видно насколько временами большая разница между этими значениями что говорит о наличии большого хвоста справа. В одном из источников приводиться следующее правило: выбросом можно считать любое наблюдение лежащее вне интервала **(анмация)**. Но в нашем случае не следовало как-либо работать с левым хвостом, потому мы немного изменили правило **(анимация)**. Так любые наблюдения вне этого интервала были исключены из выборки, описательные статистики приняли вид **(анимация)** – ситуация заметно улучшилась.

**Слайд 8**

Далее мы перешли к отбору покзателей. Для того, чтобы упростить процесс дальнейшего построения модели требовалось снизить размерность данных. В данной работе применен авторский подход к этой процедуре.

Для оценки качества классификационных моделей широко используется ROC анализ – методика которая позволяет дать характеристику на сколько отличаются распределения классов вдоль некоторой переменной.