**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 выступление описание модели подбор точки отсечения и оценка выгод от использования модели.

Соответствующие пункты вы можете найти в бумажных презентациях представленных вам.

**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.

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 исследуемых классов вдоль некоторой переменной. Обычно, ROC анализ используется уже в процессе валидации готовой модели, наша идея состояла в том, чтобы исследовать насколько успешна будет классификация в случае, если использовать всего один показатель – эти характеристики и служили основой для принятия решения о включении того или иного показателя в модель.

Вашему вниманию предложен рисунок 3, там вдоль нормированного значения некоторого показателя представлены эмпирическая функция распределения вероятности сверху и функция распределения вероятности снизу для класса 0 и класса 1. Понятно, что чем больше различаются между собой распределения этих классов, тем лучше рассматриваемый показатель подходит для классификации.

Для описания различия распределений в виде одного числа обычно используются две статистики статистика Колмогорова-Смирнова, обозначенная на рисунке (KS) и AUC. Более подробно остановимся на статистике Колмогорова-Смирнова – это по сути максимальное расстояние между функциями распределения двух классов. Мы подготовили такую визуализацию **(анимация)**, которая наглядно показывает – чем больше распределения классов различаются тем большее значение принимает статистика Колмогорова-Смирнова.

Эта статистика особенно важна для анализа такого толка потому, что за ней стоит одноименный статистический тест, нулевая гипотеза которого состоит в том, что сравниваемые распределения действительно отличаются. Опираясь на этот статистический тест мы и проводили отбор показателей в модель.

Тут справедливо обсудить как мы боролись с большим числом пропусков в данных. Основным метод выступало восприятие пропуска как отдельного уровня переменной для номинативного показателя, численные показатели мы пытались привести к номинативному, если для преобразованного показателя удавалось подтвердить разницу распределений то мы его воспринимали как номинативный, соответственно озвученное выше правило применялось к нему.

**Слайд 9**

Совокупность названных методов привела нас к следующей структуре данных. Более подробную информацию о данных, которые используются при обучении модели ищите в таблице 5.

Некоторую справку относительно количественных характеристик выборки можно получить в таблице 6 **(анимация)**. Там в первом столбце “Полные данные” указано что для моделирования после всех преобразований у нас осталось порядка 216 000 записей, из которых всего 3% дефолт. Обыденной практикой в машинной обучении является разделение выборки на тренировочную и тестовую мы делили в соотношении 3:1 и сохранением соотношения предсказываемых классов.

**Слайд 10**

Что касается обучения модели, там, по сути все строиться вокруг решения достаточно сложной задачи многомерной нелинейной оптимизации. Для обучения модели мы использовали метод обратного распространения ошибки, который представляет собой модификацию метода градиентного спуска с особым, удобным способом вычисления частных производных по коэффициентам модели.

Модель на который мы остановились, содержит всего 1 скрытый слой и 113 нейронов в нем. К сожалению представить модель в аналитической форме не представляется возможным. Для того нам бы понадобилось под сигмоиду **(анимация)** поместить линейную комбинацию из 113 ReLU функций **(анимация)** под каждой из которых лежит линейная комбинация входной информации о рассматриваемом кредитополучателе.

**Слайд 11**

Для оценки качества классификационных моделей принять использовать ROC анализ. В данном случае основным показателем качества выступает ROC кривая. Для всех предсказанных на тестовых данных вероятностях дефолта вычисляется доля ошибок предсказания дефолта (False Positive Rate; FPR) и доля правильных предсказаний дефолта (True Positive Rate; TPR). Если на промежуточных вероятностях TPR заметно опережает FPR, ROC кривая приобретает характерный для нее вогнутый вид. ROC кривая для предложенного нами классификатора представлена на рисунке 4.

Для характеристики качества классификатора одним числом используется площадь под ROC **(анимация)** кривой (Area Under Curve; AUC). Для предлагаемой модели AUC составил 0.8 что считается более чем удовлетворительным результатом.

**Слайд 12**

Отдельной задачей можно считать подбор точки отсечения – такой вероятности, что если договор получает оценку вероятности ниже то он считается не склонным к выходу в дефолт в противном случае рассматривается как потенциальный дефолт. Несколько возможных точек отсечения по строкам и соответствующие им характеристики классификатора по столбцам представлены в таблице 7.

Точка отсечения чаще всего выбирается совместно с начальством, но наше дело предложить обоснованные варианты. На рисунке 5 можете видеть сверху гистограмму частостей для двух классов вдоль предсказаний модели, а снизу эмпирическую функцию распределения двух классов.

Первый способ определения точки отсечения следующий – в точке соответствующей KS статистике **(анимация)**. Идея следующая – это точка где доля ошибочно выданных кредитов плохим кредитополучателям сильнее всего отстаёт от доли правильно выданных кредитов хорошим кредиполучателям. В нашем случае она составила 0,049. Удается правильно определить 171 с половиной тысячу хороших клиентов и 4300 клиентов которые позже выйдут в дефолт.

Так же была предпринята попытка оценить экономические выигрыши от использования данного классификатора, это блок “Экономические показатели” в таблице 7. Так как мне до сегодняшнего дня не удалось получить информацию о течении каждой задолженности я вычислял процентные выгоды и потери от невозврата в предположении что всякая задолженность протекает строго по графику, потому процентные выгоды могут быть несколько завышены а потери от невозврата занижены, ведь в случае если кредит был возвращен быстрее чем ожидалось то банк получил меньше процентов а договор вышедший в дефолт мог в начале погашаться быстрее – как следствие уменьшить сумму потерь от этого договора. Кроме того в этой оценке никак не участвует информация о не выданных кредитных договорах так как мне не удалось быстро получить информацию в нужной для новой модели форме, это может заметно занижать реальные экономические выгоды от использования новой модели так как не учитывается та ситуация понижения потерь от недовыданных кредитов.

И так если пересчитать выгоды от использования новой модели, при названной точке отсечения, то получается, что из-за строгости новой модели банк бы недополучил процентов на более 5 миллионов рублей в то время как на невозврате удалось бы сэкономить всего 697 тыс. рублей. В силу названных выше причин прошу не относиться к этой невыгоде слишком серьезно скорее всего в реальности потери на процентах были несколько ниже, но все же на тех данных, что есть использование этой модели при такой точке отсечения очевидно не рентабельно.

Мы попробовали подобрать точку отсечения так, чтобы максимизировать выгоды от использования модели **(анимация)** модель стала намного лояльнее и выдает кредит в том случае если предсказанная вероятность дефолта меньше 34 сотых. В таблице 7 это последний столбец. Тут удается предотвратить невозврат в 15 тыс. бел. руб. при том что на процентах теряется чуть более 6 тыс. выгоды от использования модели очевидно составят в районе 9 тыс. бел. рублей. Но в силу названных выше причин все-же я бы рекомендовал делать модель более строгой.