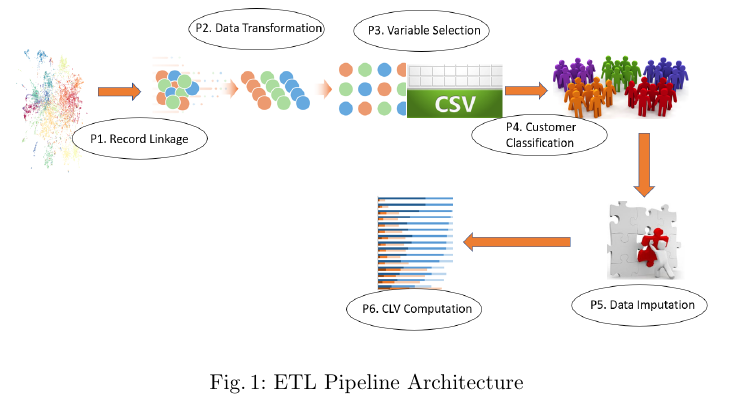
**Предсказание клиентского оттока для страховых данных**

Michael Scriney, Dongyun Nie, and Mark Roantree

Dublin City University, Ireland



**3. Преобразование данных**

В данном разделе приводится краткая схема архитектуры процессов Извлечения, Преобразования и Загрузки данных (ETL), но основной темой являются компоненты, которые новы для нашей архитектуры и являются ключевыми для восстановления (импутации) данных об оттоке. Данные, используемые в этом исследовании предоставлены нашим партнером из страхового сектора. В данной сфере сделки направлены на продажу страховых полисов, а не на создание профилей клиентов и, таким образом, этап Извлечения данных дал приблизительно 500 000 страховых полисов.

**3.1 Архитектура системы**

Конвейер ETL включает в себя набор компонентов, извлекающих данные из источников, преобразующих эти данные, чтобы соответствовать системной модели и компонентов, загружающих данные в витрину данных (куб данных) для отчетности и аналитики. Наша архитектура, показанная на рисунке 1, это особая форма ETL-конвейера [14], обоснованная определенными требованиями поставленной задачи (время жизни клиента) и природой данных. В частности, в данной работе используется датасет, основанный на понятии полиса, а не клиента. Из-за чего он не подходит для анализа клиентов. Поэтому первым шагом стало связывание документов, при котором после сбора данные приводились к формату клиентов, где в каждом документе клиента указан один или более страховой полис. Данное исследование было представлено в [10] и, хотя в результате были получены более целостные данные о клиентах, они не подходили для алгоритмов восстановления, использующихся для восстановления недостающих переменных времени жизни клиента (CLV — Customer Lifetime Value). Также датасет все ещё не был классифицирован по типам клиентов (плохой, хороший, средний).

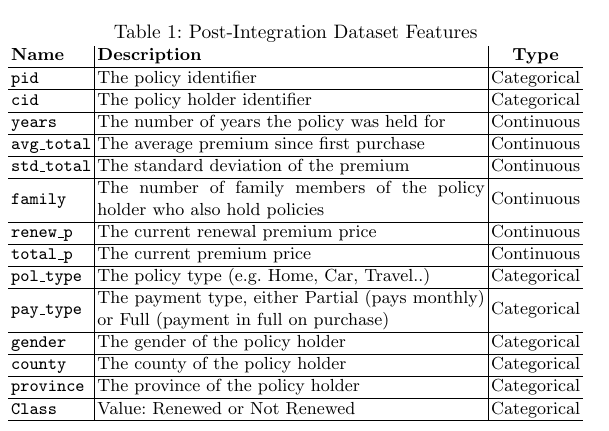
**3.2 Преобразование данных для анализа оттока**

В данном исследовании работа была сконцентрирована на пунктах P2 и P5, показанных на рисунке 1. Используемые данные изначально основаны на двух крупных выборках: детализации и агрегата. Детализация показывает данные относительно полисов по годам, включая в себя тип, текущие взносы и взносы за продление по каждому полису. Агрегат это группирование данных в разрезе клиентов по единым документам клиентов, владеющими страховыми полисами. Он был получен в результате предыдущей работы [10]. Эти крупные выборки объединены с другими источниками данных из хранилища данных для создания датасета, подходящего для предсказания клиентского оттока. Существует три процесса, входящих в преобразование (P2) датасета, подходящего для анализа оттока: агрегация (группирование), аугментация (раздутие) и предобработка. В результате агрегации определяется информация о продлениях по каждому полису. Аугментация добавляет различные свойства к датасету, например информацию о клиенте и ценах. Эти два процесса можно отнести к процессу Извлечения (Extract) и процессу Преобразования (Transform) в типовом ETL-конвейре. Последний процесс Подготовки выполняет последние преобразования над датасетом, чтобы подготовить его к алгоритмам машинного обучения.

**Агрегация.** Цель первого шага заключается в создании ориентированного на страховые полисы представления данных, содержащего те полисы, которые могут или не могут быть продлены. Для этого используется операция RollUp на детализированных данных, чтобы создать агрегат, состоящий из идентификатора полиса (id полиса), количества лет, в течение которых полис активен (years held) и был ли полис продлен (renewed). Итого, датасет, использованный в данной работе содержит 443 893 уникальных полиса, из которых 300 646 не были продлены и 143 247 были продлены клиентом.

**Аугментация.** Следующим шагом является аугментация, при которой данные из хранилища данных интегрируются с агрегацией страховых полисов. В сумме семь дополнительных параметров были интегрированы: цены полисов, владельцы семейных страховых полисов, последний взнос за продление, вид страхования, местоположение, способ оплаты и пол. Цены полисов состоят из среднего взноса и стандартного отклонения взноса, который может отражать количество изменений стоимости полиса по годам. Владельцы семейных страховок это число членов семьи, приходящееся на клиента, который также владеет страховым полисом компании. Последний взнос за продление это последний взнос за определенный полис. Вид страхования это один из 4 типов страховки: частное страхование авто, страхование авто юр. лиц, недвижимости и туристическая страховка. Местоположение — страна проживания клиента. Метод оплаты отражает производится ли оплата единоразово или помесячно. И, наконец, пол относится к полу владельца полиса. Результатом является датасет с 14 измерениями, включая классификацию по продлен / не продлен, как показано в таблице 1, где Name — имя признака, Description — краткое описание признака и Type разделяет качественные и количественные параметры. Для нашей цели измерения уникальных идентификаторов (pid and cid) не были использованы.

**Предобработка.** На этапе предобработке выполняются 4 шага: очистка, сэмплирование (выборка), кодирование и разбиение. Из датасета были удалены только 27 записей, в датасете осталось 443 866 строк. Определенить продлен ли полис фактически является задачей классификации. Наименования классов для каждого полиса — Продлен или Не продлен. Как обычно при работе с реальными данными, у полученного датасета наблюдалась несбалансированность классов, при которой 300 621 запись с классом «Продлен», а остальные 143 245 с классом «Не продлен». Эта несбалансированность классов может существенно повлиять на результаты классификации, но есть 3 метода как правило используемых для решения этой проблемы: уменьшение мажоритарной выборки (undersampling), увеличение миноритарной выборки (oversampling) и синтетические выборки. Так как уже имеется большое количество записей с классом «Продлен», был выбран метод уменьшения мажоритарной выборки. С применением этого метода было случайно выбрано 143 245 записей с признаком «Не продлен», чтобы уравнять кардинальность обоих классов. Обратной стороной данного подхода является то, что некоторые данные с классом «Не продлен» могли повысить эффективность анализа. Этому посвящена часть рассуждений в заключении. После уменьшения мажоритарной выборки датасет составил 286 490 записей с равным распределением классов «продлен» и «не продлен» (143 245 записей каждый). Этап кодирования преобразует качественные измерения, подготавливая их к алгоритмам машинного обучения. Измерения, подвергшиеся кодированию это вид страхования, тип оплаты и местоположение. Последний этап разделяет данные на тренировочный и тестовый набор данных, используя принцип 80/20.



**4. Выбор алгоритма и проверка.**

Из-за особенностей страховых данных и того факта, что исследование времени жизни клиента носит достаточно теоретический характер было решено использовать набор статистических методов и пытаться определить, какой из них лучше работает. В данном разделе начнем с набора алгоритмов для восстановления оттока, затем продолжим обсуждение оценкой методов и результатов и, наконец, обсудим результаты.

**Выбор алгоритма.** Процесс определения клиентского оттока в любой области является задачей классификации с двумя классами: «Продлен» и «Не продлен». Были применены следующие методы классификации: Наивный байесовский классификатор; полиномиальный наивный байесовский; два вида метода опорных векторов; два дерева решений; и набор нейросетевых конфигураций. Метод опорных векторов чаще всего используется в классификации. Для проведения экспериментов был использован линейный опорный вектор для создания основы для других методов. Две экспериментальных конфигурации с использованием наивного байесовского классификатора были применены, один с использованием модели Бернулли и второй с применением полиномиальной модели, которая продемонстрировала повышение производительности на двоичных данных в некоторых примерах [9]. Для обеих моделей 100 различных альфа значений были использованы для каждого, изменяя свое значение от 0.0 до 1 с шагом 0.01. Два дерева решений, использующих алгоритм CART (дерево классификации и регрессии) были применены, первое с использованием энтропии в качестве разделительной меры и второе с примесью Джини. Для обоих подходов дерево решений было построено для каждой глубины до того момента, пока не будет достигнута максимальная глубина и на каждой глубине тестовые данные были использованы для вычисления точности дерева. Искусственные нейронные сети (ИНС) широко применялись в предсказании клиентского оттока из-за их возможности моделирования взаимодействия между признаками, которые в противном случае могли остаться незамеченными. Для проведения экспериментов 20 различных конфигураций ИНС было создано с различной настройкой гиперпараметров.

**Вычисляемые метрики.** Опишем оцениваемые метрики для сравнения различных моделей предсказания. Меры TP, TN, FP и FN это истинно положительная, ложно положительная, истинно отрицательная и ложно отрицательная соответственно. Параметры: общая точность модели, точность TP, полнота, специфичность, F1-мера. Общая точность (процент правильных ответов) неэффективна при оценке классификатора, но предоставляет хорошую основу. Точность, полнота и специфичность предоставляют больше информации о фактическом качестве работы классификатора в пределах классов. Все конфигурации моделей будут проверены с использованием всех 5 метрик.

**Результаты и обсуждение.** Начнем данный раздел с обзора 4 различных алгоритмов по отдельности, приводя оценку качества их работы. Затем сравним все алгоритмы с использованием различных конфигураций для более сложных моделей. Как и ожидалось, метод линейных опорных векторов показал свою малую эффективность с общей точностью 0,754 и F1-метрикой 0,76. Тем не менее, данная модель всегда рассматривалась как основа для оценки других моделей. Для обеих моделей наивного байеса 100 различных альфа значений были использованы, изменяясь от 0.0 до 1 с шагом 0.01. Примечательно, что эти изменения не повлияли на общую точность среди конфигураций модели. У алгоритма с использованием энтропии наиболее качественным оказалось дерево с глубиной решения 11 и общей точностью 88,82%. У дерева с примесью Джини самое высокое качество также было достигнуто при глубине 11 и очень похожей точности 88,72%. Результаты топ-5 самых эффективных конфигураций приведены в таблице 2, где id — идентификатор эксперимента, epoch — количество эпох, hlayer — количество скрытых слоев, hnode — количество скрытых узлов, tr\_ac — общая точность на данных для обучения, te\_ac — точность на тестовых данных, и, наконец, te\_l — потери на тестовых данных. Для всех 5 конфигураций был использован процент исключения 0,02. Исходя из данных таблицы 2 эксперимент 5 показал наибольшую эффективность с общей точность 0,888 на тестовых данных. Модель состояла из одного скрытого слоя с процентом исключения 0.2%. Также были другие модели с повышенной точностью на тренировочных данных, но они не были включены в таблицу, так как оценка точности на тестовых данных составила менее 88,6%, что как правило говорит о переобучении.

