Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра иностранных языков технических факультетов

ПИСЬМЕННЫЙ ПЕРЕВОД

по дисциплине «Иностранный язык»

Тема: Predicting Customer Churn for Insurance Data

Рецензия:_____

_____

Выполнил:

Студент: Сухих А.С.

Группа: ПММ-21

_____
подпись

«_____»_____20___г.

Проверил:

Преподаватель: Балобанова А.Г.

Балл: _____ , ECTS_____

Оценка _____

_____
подпись

«____»_____20___г.

Новосибирск 2022

Predicting Customer Churn for Insurance Data

Michael Scriney, Dongyun Nie, and Mark Roantree

Dublin City University, Ireland

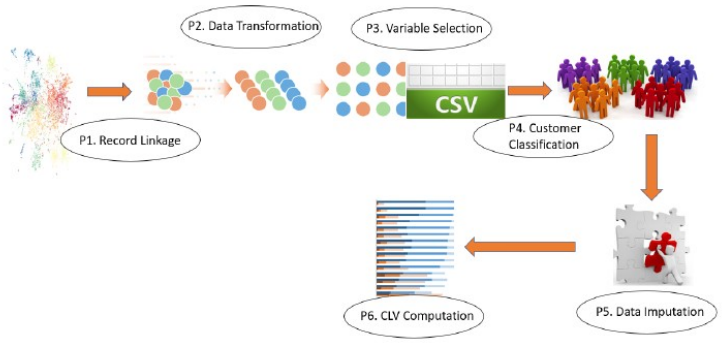

Fig. 1: ETL Pipeline Architecture

3. Data Transformations

In this section, we provide a brief outline of the Extract Transform Load (ETL) architecture but focus on those components which are novel to our architecture and crucial to imputing retention data. The dataset used in this work originated from our collaborator in the insurance sector. In this sector, transactions focused on selling policies and not on building customer profiles and thus, the Extract component in our architecture acquired approx. 500,000 insurance policies.

3.1 System Architecture

An ETL pipeline comprises a series of components extracting data from input sources, transforming the data to match the system's data model, and loading into a data mart (data cube) for reporting and analysis. Our approach, shown in figure 1, is a specialised form of ETL [14], due to the specific requirements of the task (customer lifetime value) and the nature of the data. In particular, this work began with a dataset that was policy-focused and not customer-focused. In effect, it was not suited to analysis by customer. Thus, the first step involved a process known as record linkage where, upon acquisition, data was pivoted to be customer-focused, where a customer record contained 1 or more policies. This work was presented in [10] and, while it provided a more holistic customer record, the data was not suited to the imputation

algorithms necessary to impute the missing CLV variables. In addition, the dataset was still unclassified in terms of customer types (good, bad, average).

3.2 Churn Analysis Data Transformation

In this paper, we focus on components P2 and P5 from figure 1. The data used is initially based on two large imports: detail and aggregate. Detail provides a policy centric view year-on-year, recording the type, current and renewal premium for each policy. Aggregate is an aggregation of a unified customer record, detailing high level information on customers who hold policies generated during earlier work [10]. The large data imports are combined with other data sources within the warehouse to provide a dataset suitable for predicting customer churn. There are three processes involved in the transformation (P2) of a dataset suitable for churn analysis: Aggregation, Augmentation and Preparation. Aggregation constructs the initial per-policy view which provides information on policy renewals. Augmentation adds features to this dataset such as customer information and pricing. These two processes can be equated to the E and T processes within a standard ETL (Extract, Transform, Load) architecture. The final process Preparation provides a final transformation of the dataset so that is it ready for machine learning algorithms.

Предсказание клиентского оттока для страховых данных

Майкл Скрини, Дон-Юн Ни и Марк Раунтри
Городской Университет Дублина, Ирландия

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.

Список литературы

Scriney M., Nie D., Roantree M. Predicting customer churn for insurance data //International Conference on Big Data Analytics and Knowledge Discovery. – Springer, Cham, 2020. – c. 259-260.