Научная статья
Original article

Интеллектуальный анализ статистических данных трафика электронного документооборота

Ткаченко А.Л.
Калужский государственный университет им. К.Э. Циолковского, Калуга, Россия
Автор-корреспондент: tkachenkoal@tksu.ru

Аннотация: Востребованность и растущий интерес к программным продуктам становятся отправной точкой для совершенствования бизнес аналитики, потому что основным содержанием этого процесса является анализ большого количества данных, а программный продукт позволяет автоматизировать этот процесс. Инструментами бизнес-аналитики являются аналитические системы, позволяющие обработать, структурировать и предоставить исходную информацию таким образом, чтобы она была удобна для оперирования в принятии управленческих решений, ведь без качественной бизнес-аналитики сегодня практически невозможно обойтись. В данной статье рассматривается анализ статистических данных с помощью аналитической платформы Loginom. Имея исходный набор данных по префиксам различных документов, необходимо привести их к удобному для пользователя виду и исключить объекты, используемые наименее часто. Исходные данные были отформатированы, отфильтрованы и отсортированы, благодаря чему, приведены к компактному и удобному виду. Такой инструментарий позволяет исключать ненужные данные или уменьшить количество рассматриваемых объектов при большом объеме рассматриваемой выборки, что приведёт к компактной таблице, содержащей только требуемые для пользователя данные из всех исходных, благодаря чему будет проще оперировать только нужными данными.

Ключевые слова: анализ данных, статистический анализ, электронный документооборот, аналитическая платформа, Low code, Big Data, Loginom.

Для цитирования: Ткаченко А.Л. Интеллектуальный анализ статистических данных трафика электронного документооборота. 2022. Т.2, №4, с. 6-17

Intelligent analysis of statistical data of electronic document management traffic

Tkachenko A.L.
Kaluga State University K.E. Tsiolkovsky, Kaluga, Russia
Corresponding author: *tkachenkoal@tksu.ru*

Abstract: Demand and growing interest in software products are becoming the starting point for improving business intelligence, because the main content of this process is the analysis of large amounts of data, and the software product allows you to automate this process. Business intelligence tools are analytical systems that allow you to process, structure and provide source information in such a way that it is convenient for operating in making management decisions, because today it is almost impossible to do without high-quality business intelligence. This article discusses the analysis of statistical data using the analytical platform Loginom. Having an initial set of data on the prefixes of various documents, it is necessary to bring them to a user-friendly form and exclude the objects that are used least often. The source data has been formatted, filtered and sorted, thanks to which it has been reduced to a compact and convenient form. Such a toolkit allows you to exclude unnecessary data or reduce the number of objects under consideration with a large volume of the

sample under consideration, which will lead to a compact table containing only the data required for the user from all initial ones, which makes it easier to operate only with the necessary data..

Keywords: data analysis, statistical analysis, electronic document management, analytical platform, Low code, Big Data, Loginom.

For citation: Tkachenko A.L. Intelligent analysis of statistical data of electronic document management traffic. Smart Digital Economy. 2022. Vol.2, №4, pp. 6-17

Введение

Объемы данных, которые необходимо накапливать и анализировать, увеличивается с каждым днем. По словам аналитиков, в 2025 году, объем этих данных вырастет в 5 или более раз. И в связи с таким огромным объемом информации и данных, используемое во многих компаниях программное обеспечение для анализа не справляется.

Если посмотреть исследования, которые проводили TechValidate, почти 50 процентов опрошенных считали, что программное обеспечение уже достигло пределов своих возможностей и нет смысла его дорабатывать, ему уже некуда развиваться, оно просто не справится. Ну и 70 процентов респондентов ответили, что может развиваться и есть куда, но задаваемые запросы обрабатываются слишком медленно или просто вызывают зависания и сбои программного обеспечения. Из-за таких условий у многих встают несколько достаточно важных вопросов:

- Если стандартный MS Excel уже достиг своего предела, как заниматься анализом новой информации?
- Как и где можно научиться профессии Data Scienc?
- Где найти и выбрать более подходящий программный продукт, который стравится с новыми объемами данных?
- Как понять, насколько предполагаемые сценарии и гипотезы пригодны для реализации в жизни?

Аналитическая платформа Loginom может помочь в решении данных вопросов. Сейчас наблюдается большое количество вакансий для программистов, эта сфера расширяется быстрыми темпами, но на рынке труда требуемые специалисты как раз уменьшаются. Многие уже нашли себе работу и не спешат куда-то уходить. Либо же работают на себя, что тоже не дает прироста в работниках в компаниях. В то же время, для решения данной проблемы появилась концепция low-code. Её особенностями являются уменьшение количества кода, если не его отсутствие, при этом способность выстроить различные сложные процессы со встроенными в них логикой. Ещё одним преимуществом является снижение порога входа в данную сферу для специалистов. Это повышает процент тех, кто может начать работать здесь, и уменьшается время, затраченное на изучение рабочих процессов. И самым важным является экономия времени, то есть его сокращение, на разработку того или иного сценария. При чем эта экономия может составить от 40 до 90 процентов.

Платформа Loginom использует концепцию low-code. Это позволяет сделать продвинутую аналитику более доступной для бизнес-пользователей. Все процессы анализа имеют виртуальный конструктор, то есть каждый из них можно настроить визуально, будь то

интеграция, подготовка данных, моделирование или визуализация. Время от тестирования гипотезы до создания уже рабочего процесса в жизни сокращается примерно на 60 процентов и не требует особых знаний в плане кода. Большое количество различной информации есть у не менее большого количества компаний, это делать они научились. Но выносить и извлекать из этой информации важные и ценные знания могут позволить себе далеко не многие. Для того, чтобы использовать алгоритмы машинного обучения, которые основаны на математике и статистике, специалист должен быть знаком с основами анализа данных, понимать теорию вероятности и разбираться в базах данных. Также должен уметь использовать различные методы в решении той или иной проблемы и понимать, какое решение подойдет, а какое нет.

У Loginom есть курсы Loginom e-Learning, в которых они сконцентрировали всю основную информацию для того, что человек быстро погрузился и освоил основы Data Science и продвинутой аналитики. Для практики можно использовать бесплатную версию платформы Loginom Community Edition, в ней рекомендуется выполнять задания курса, то есть при наличии желания и времени, человек может моментально получить результат обучения, основываясь на своих собственных данных.

При работе с анализом больших данных, то есть Big Data, с которыми сталкиваются при решении бизнес-задач, используют огромное количество разных инструментов. Каждый из таких инструментов обладает различной успешностью и достоверностью обработки, моделирования и визуализации имеющихся данных. И многие часто полагаются лишь на маркетинговое описание продукта. В то же время Loginom имеет бесплатную версию, на которой можно самому протестировать работу с реальными данными.

Аналитик может свободно поработать с данным продуктов, сравнить все параметры его работы с альтернативами, будь то скорость, качество анализа, удобство работы с интерфейсом и подобное. Благодаря этому человек может сделать свой обоснованный выбор. Для некоммерческого использования любому доступна бесплатная версия.

Low code платформы по прогнозам агентства аналитики International Data Corporation разовьются и будут использоваться для создания чуть ли не большинства всех новых решений в любой сфере, так как они существенно сокращают любые затраты времени на каждом этапе, будь то создание прототипа, разработка, тестирование сценария или гипотезы, или же развертывание уже готового решения.

Возможности программного обеспечения от Loginom рекомендуются для проектов, которые используют Agile-методологию. При данном подходе объединяются различные способы выполнения проектов основных на гибких методологиях. Этот метод позволяет быстро получить прототип, который уже готов к работе, его останется только сформулировать работу будущего бизнес-процесса. Один из принципов этой методологии – «простота – искусство не делать лишней работы». Это относится и к проектам с анализом данных. Важно сразу понять, ещё до построения модели, каким будет качество данных, которые уже имеются. Часто оказывается так, что они для анализа совершенно не подходят. Для понимания, какие данные пригодны к анализу, а какие нет, в Loginom есть компонент, которые делает это автоматически. Это упрощает аналитику на первых этапах. Для данного компонента необходимо задать входные и выходные наборы полей, после чего функция сама выявит полные дубли и противоречия в наборе. Дублями считаются строки, с полным совпадением

полей, которые сравниваются. У платформы Loginom есть функционал, позволяющий тестировать различные гипотезы и идеи без детального технического задания, лишь с помощью визуального проектирования. После, эти прототипы уже можно будет использовать для опоры при построении будущих проектов, что, естественно, уменьшит сроки и стоимость введения решений в жизнь.

В целом работа в данной платформе состоит из пяти этапов. На первом происходит извлечение данных из файлов, баз данных, веб-сервисов или бизнес-приложений. Данная информация изучается и использует программный анализ данных, чтобы иметь возможность очистить, провести предобработку и подготовку к дальнейшим действиям, связать какую-то информацию между собой и систематизировать её. Далее происходит отбор по конкретным факторам, используется машинное обучение, которое легко настраивается. Также используется прогнозирование и различные сложные расчеты, как встроенные в платформу, так и новые, созданные пользователями. После обработки, подготовки и основного анализа информации, программное обеспечение позволяет визуализировать полученные данные, провести ещё ряд многомерных анализов и используя новые результаты подвести итоги и интерпретировать их для понимания человеком. И на основании данных итогов принимается решение о введении данного сценария в жизнь. При положительном решении происходит интеграция в IT-инфраструктуру организации, либо же загрузка в базы, в файлы или в приложения. Также итогом решения может быть публикация спроектированного веб-сервиса.

Платформа Loginom предоставляет широкий выбор, какую информацию и из какого источника загрузить для обработки. Она позволяет подключиться сразу к нескольким источникам и приемникам данных и настроить ETL-процессы. В пример источников и приемников информации можно привести различные базы данных - Oracle, MS SQL, PostgreSQL, ClickHouse, BigQuery, используемые в другом ПО файлы - Excel, CSV, XML, Loginom Data File. Такая интеграция с веб-сервисами и возможность опубликовать свои разработки значительно упрощает интеграцию любой компании.

Также для тех, кто всё же использует программирование кодом в своей работе можно выделить возможность пользоваться им в данном программном обеспечении. В последних версиях была добавлена поддержка Python. То есть в данное время можно встраивать в сценарии расчеты, написанные на этом языке программирования. Присутствует поддержка нескольких популярных python библиотек, таких как NumPy, Pandas, Scikit-learn и другие.

Стоит упомянуть встроенные возможности визуализации данной платформы. Она поддерживает десятки различных возможностей для визуализации не только небольших отчетов, но и больших наборов данных, а также можно получить панель различных отчетов для получения результатов обработки. Данные могут визуализироваться в режиме реального времени, то есть «на лету». Многомерные данные можно сортировать как требуется, группировать по каким-либо признакам, фильтровать отдельные значения и строить диаграммы, связанные с кросс-таблицей. Отдельной функцией можно назвать возможность детализации по отдельной ячейке, что часто бывает требуется. Визуализация может быть, как по системе OLAP-куб, так и используя табличные данные, имеет возможность использовать специализированные визуализаторы, чтобы оценить качество модели и интерпретировать результаты.

Ещё одной положительной чертой платформы Loginom является, что она не ограничивается настольной версией Loginom Community Edition, а имеет серверную редакцию. В ней доступны дополнительные инструменты для работы в коллективе, можно также обрабатывать данные в пакетном режиме, вызов различных, как своих, так и сторонних веб-сервисов.

Работа в коллективе оптимизирована в последних версиях, так как раньше сервер со сценарием мог выдавать ошибки о том, что пакет уже занят другим пользователем. При этом было не совсем понятно, какой именно пользователь или процесс занимает пакет. Это проводило к трудностям и лишним действиям в Loginom Integrator. Сейчас же добавлен диспетчер открытых пакетов и сессий, где можно полностью управлять ими, например приостановить выполнение пакета или же сбросить сессию пользователя.

Альтернативой также можно назвать аренду платформы в Яндекс.Облаке на любой срок для чего угодно, будь то развернуть уже готовое решение, или же тестировать и экспериментировать. Работа в облаке позволяет обеспечить отказоустойчивую инфраструктуру для развертывания боевых решений.

Основным содержанием бизнес аналитики является сбор и анализ большого количества данных, что помогает принимать более эффективные управленческие решения. На текущий момент на рынке представлен большой ряд программных продуктов, имеющих схожий между собой функционал. Loginom – аналитическая платформа, замечательно подходящая для решения возникающих аналитических проблем. Программа позволяет произвести изучение, сортировку, замену и фильтрацию данных, построить прогнозируемый в будущем исход, также обладая множеством других полезных функций [7].

Сценарий в Loginom представлен в форме дерева, состоящее из узлов-обработчиков данных и визуализаторов. Этот метод удобен, когда нужно реализовать простую логику. Однако у этого метода имеются свои минусы: постоянно необходимо объединять, разделять, подтягивать данные на различных этапах анализа, декомпозировать большие задачи и объединять логические блоки в подмодели.

Разработка сценариев в Loginom реализовано по модели «снизу вверх», то есть необходимо наличие данных на входе.

К преимуществам данной модели можно отнести:

- высокую скорость реализации, благодаря наличию информации о структуре данных на входе;
- простоту поиска ошибок и отладки.

Недостатками модели являются:

- ориентированность на отдельных задачах;
- сложность повторного использования в аналогичных задачах;
- необходимость редактирования всего сценария при изменении входных данных.

В Loginom администрирование позволяет управлять пользователями, рабочими папками их правами доступа, а также параметрами работы сервера. Возможные варианты визуализации данных в Loginom: таблица, статистика, диаграмма, куб, матрица корреляции, факторный анализ, конечные классы, граф нейросети, дерево решений, карта Кохонена, отчет по регрессии, связи кластеров, метаданные.

В Loginom интеллектуальный анализ данных позволяет строить следующие:

- решающие деревья – логический алгоритм классификации, основанный на поиске конъюнктивных закономерностей;
- самоорганизующиеся карты Кохенена – самообучающаяся нейронная сеть без учителя, выполняющая задачу визуализации и кластеризации;
- многослойные нейронные сети - нейронные сети, в которых нейроны сгруппированы в слои. В этом случае каждый нейрон предыдущего слоя связан со всеми нейронами следующего слоя, и между нейронами внутри слоев нет никаких связей, представлены на рисунке 1.

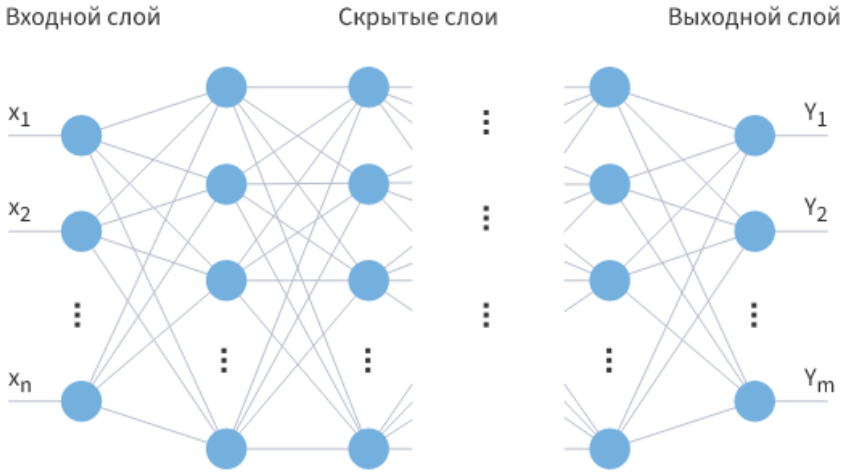

Рисунок 1. Многослойная нейронная сеть

Машинное обучение позволяет разрабатывать и строить аналитические модели, которые способны автоматически находить в данных скрытые закономерности, а также самостоятельно обнаруживать свойства, необходимые для определения этих закономерностей.

Создание собственных компонентов и подключаемые пакеты позволяет аналитику создавать собственные компоненты и размещать их в общей палитре, при этом доступ к ним определяет сам автор.

Результаты

Рассмотрим данные, полученные у компании ООО «Альтера» в которых представлены показатели статистических данных по трафику электронного документооборота, собранные в период с 2019 по 2021 год. Произведем статистический анализ, взяв для него 14 наиболее часто используемых префиксов наименований счетов фактуры в вышеуказанные года. Статистика по использованию данных префиксов приведена на рисунке 2.

КНД	Префикс	Всего 2019	В роуминге 2019	Всего 2020	В роуминге 2020	Всего 2021	В роуминге 2021
1115131	ON_NSCHFDOPPR	131481	72600	4253914	3273273	6902601	5575114
1115132	ON_NSCHFDOPPOK	69910	49327	2709964	2062617	5005624	4020111
1115125	ON_SCHFDOPPR	1197013	749498	24848	17394	2042	1266
1115126	ON_SCHFDOPPOK	617484	542917	33001	22049	2324	1709
1115127	ON_KORSCHFDOPPR	13523	7542	49168	36403	101555	65168
1115128	ON_KORSCHFDOPPOK	7849	7203	28662	20015	73319	48246
1115133	ON_NKORSCHFDOPPR	0	0	2072	123	16013	4291
1115134	ON_NKORSCHFDOPPOK	0	0	155	6	7221	1709
1175012	DP_REZRUISP	312985	150098	689350	448883	803181	601371
1175013	DP_REZRUZAK	223755	196138	521513	352874	631289	481099
1175010	DP_TOVTORGPR	130611	62987	247954	163664	334626	241169
1175011	DP_TOVTORGPOK	100878	90797	187228	122113	251932	179689
1175014	DP_PRIRASXPRIN	3	0	321	83	797	585
1175015	DP_PRIRASXSVED	0	0	0	0	0	0

Рисунок 2. Статистика документооборота ООО «Альтера»

Данные, представленные на рисунке 2, дают возможность просмотреть, оценить и проанализировать статистику использования различных префиксов благодаря их импорту в Loginom.

Также были выведены графики, наглядно показывающие количество использованных префиксов в конкретные года. Каждому типу документации соответствует две колонки, левая из которых – весь документооборот по данному префиксу, а правая – внешний документооборот. Графики за года 2019-2021 приведены на рисунках 3,4,5.

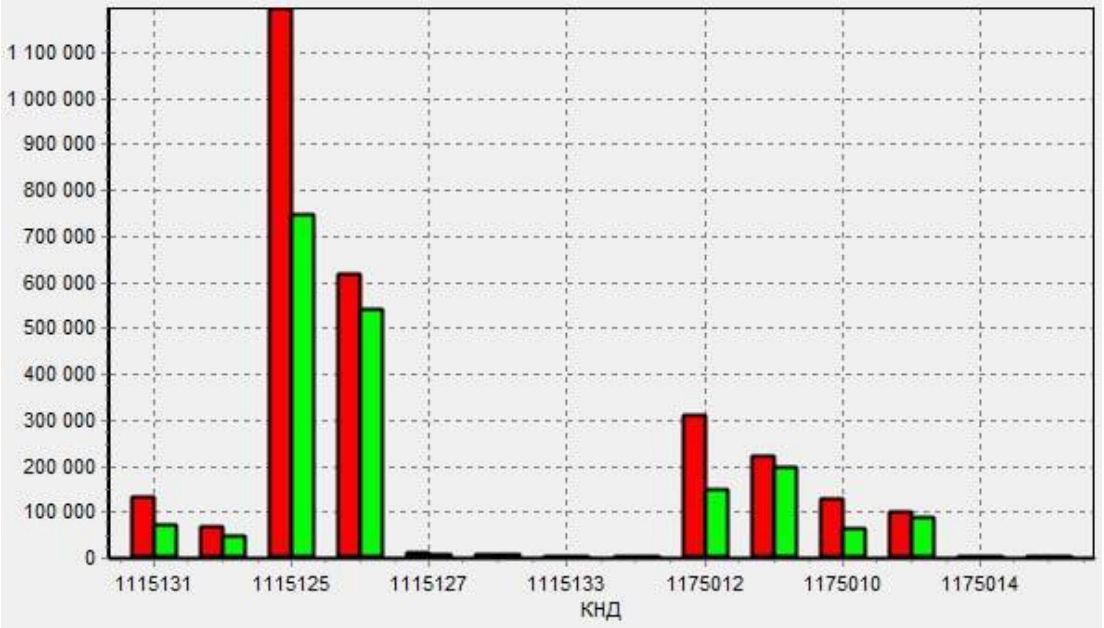

Рисунок 3. График электронного документооборота ООО «Альтера» за 2019 г.

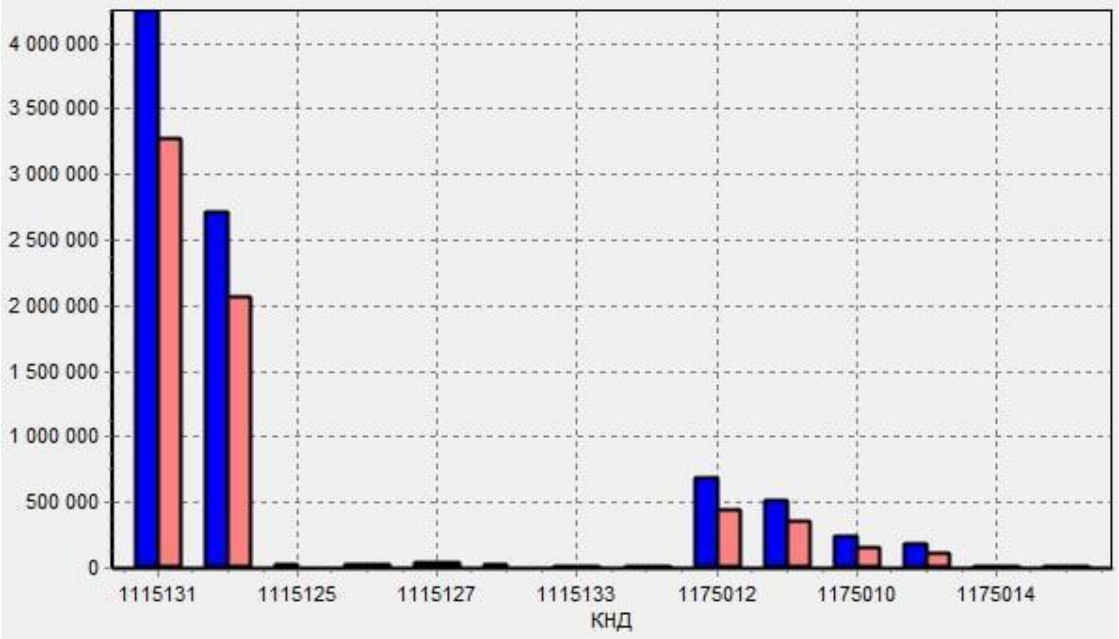

Рисунок 4. График электронного документооборота ООО «Альтера» за 2020 г.

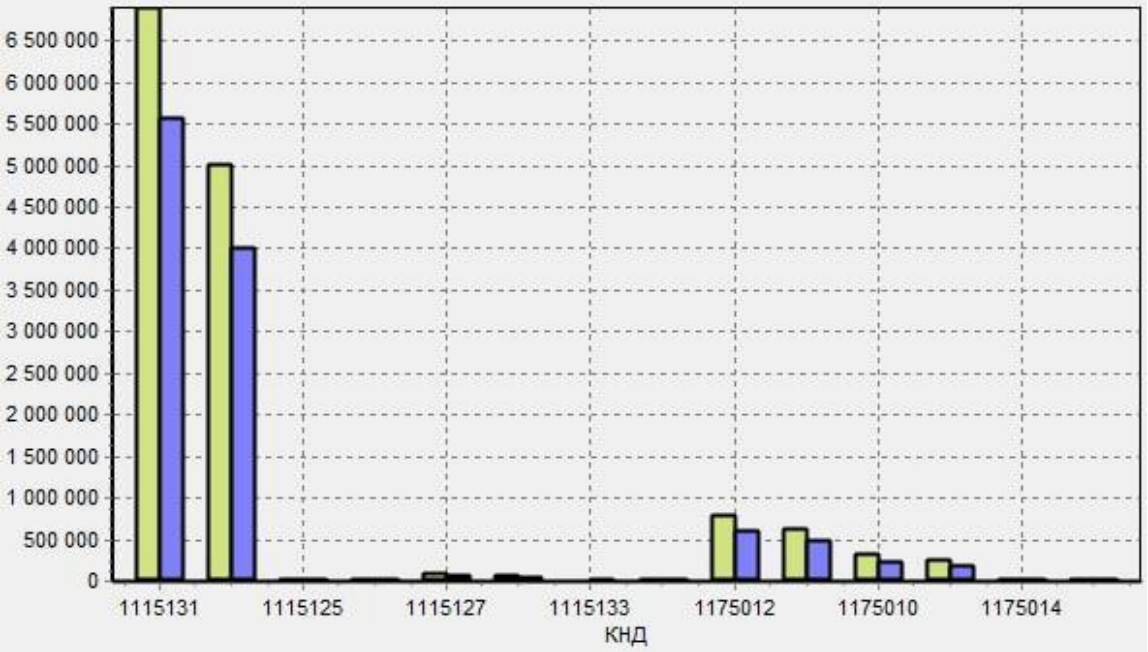

Рисунок 5. График электронного документооборота ООО «Альтера» за 2021 г.

Благодаря произведенной фильтрации данных, была произведена сортировка документации за различные года, в результате которой в каждый год были выделены три наиболее часто используемых типа документации [1-5]. Остальные значения были убраны из выборки. Изменённые графики представлены на рисунках 6,7,8.

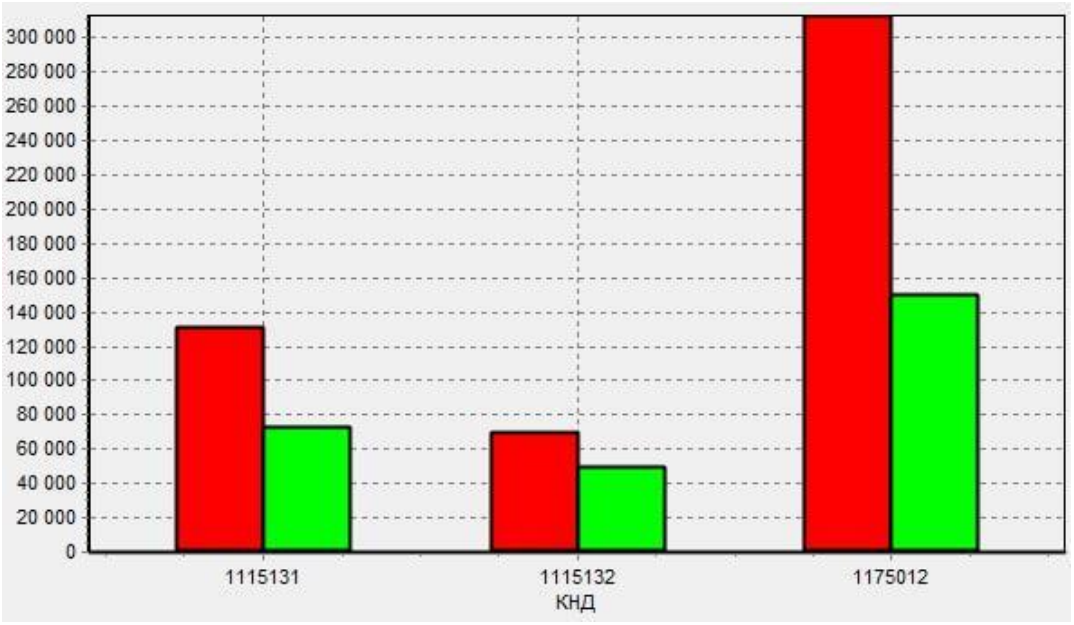

Рисунок 6. График наиболее популярных префиксов электронного документооборота ООО «Альтера» за 2019 г.

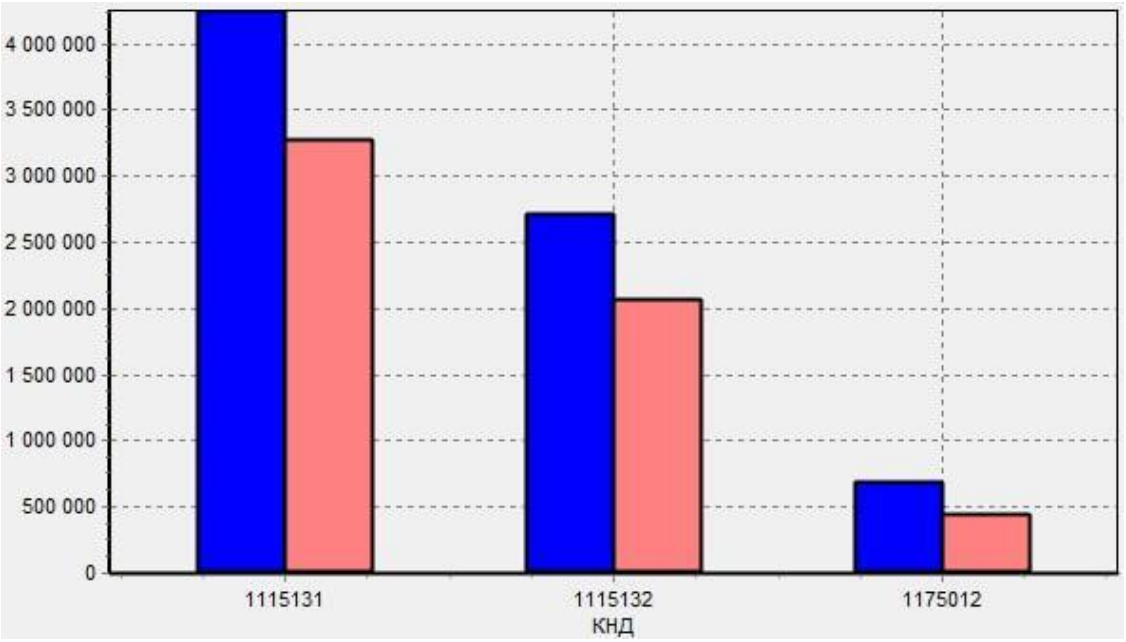

Рисунок 7. График наиболее популярных префиксов электронного документооборота ООО «Альтера» за 2020 г.

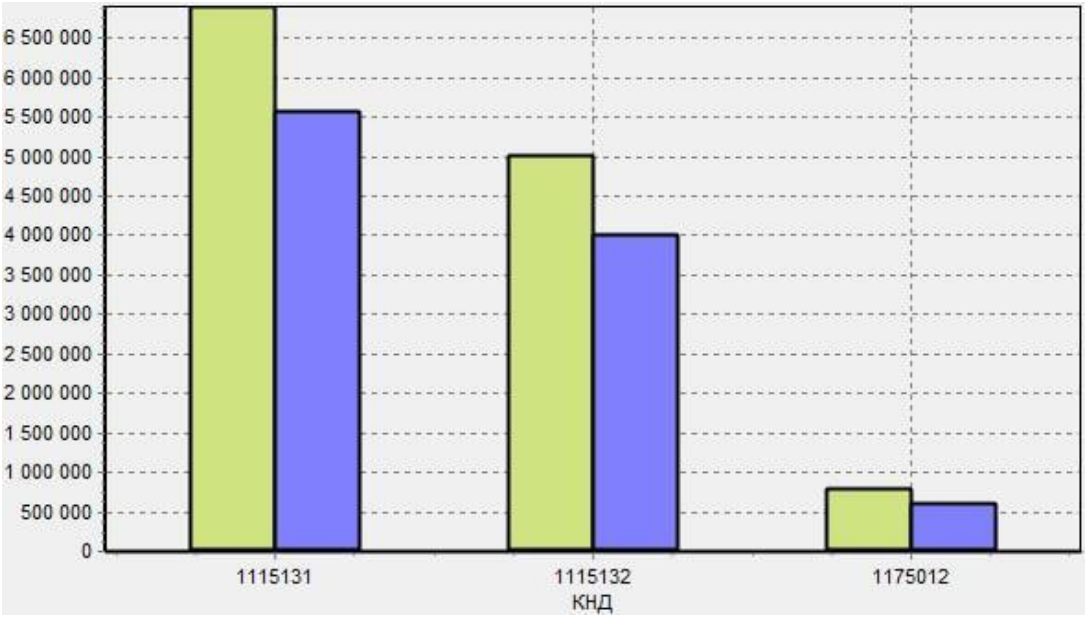

Рисунок 8. График наиболее популярных префиксов электронного документооборота ООО «Альтера» за 2022 г.

Таким образом, благодаря Loginom, набор исходных данных, представленных на таблице, был проработан, отсортирован и доведён до того вида, который может быть необходим. Отфильтрованные до удобного для эксплуатации вида представлены на рисунке 9.

КНД	Префикс	Всего 2019	В роуминге 2019	Всего 2020	В роуминге 2020	Всего 2021	В роуминге 2021
1115131	ON_NSCHFDOPPR	131481	72600	4253914	3273273	6902601	5575114
1115132	ON_NSCHFDOPPOK	69910	49327	2709964	2062617	5005624	4020111
1175012	DP_REZRUISP	312985	150098	689350	448883	803181	601371

Рисунок 9 - Доработанная таблица с исключенными данными по статистике документооборота ООО «Альтера»

Заключение

Таким образом, благодаря Loginom, набор исходных данных был проработан и доведён до того вида, который необходим. В получившемся эксперименте, с помощью имеющегося инструментария, исходная таблица со статистическими данными была отсортирована и отфильтрована до необходимого нам вида. Такой инструментарий позволяет исключать ненужные данные или уменьшить количество рассматриваемых объектов, что приведёт к компактной таблице, содержащей только требуемые для пользователя данные из всех исходных, благодаря чему будет проще оперировать только нужными данными. Loginom позволяет работать с данными, форматировать и анализировать их как требуется пользователю, что придаёт данной аналитической платформе высокий уровень значимости и востребованности по сравнению с подобными разработками.

Список литературы

1. Tkachenko, A., Lavrentev, D., Denisenko, M., Kuznetsova, V. E3S Web of Conferences, 2021, 270, 01003. https://doi.org/10.1051/e3sconf/202127001003
2. Liying, Cao, Xiaohui, San, Yueling, Zhao, Guifen, Chen *Mathematical and Computer Modeling*, 58, pp. 507-513. (2013)
3. Tarik, R. We Create a Neural Network: Trnsl. from English, 272. (2017)
4. Chesalin, A.N., Grodzensky, S.Ya., Van, T.F., Nilov, M.Yu., Agafonov, A.N. *Russ. Technol. J.*, 8 (6), pp. 167-183 (2020)
5. Bondarenko, G.G., Andreev, V.V., Stolyarov, A.A., Chukhraev, I.V., Tkachenko, A.L. Fizika i Khimiya Obrabotki Materialov, 2001, (4), pp. 94–99
6. Kashirina, I.L., Demchenko, M.V.p. 10 VSU. (2018)
7. Rashka, S.: Python and machine learning. Translator: Logunov, A.V., Movchan, D.A. (eds.), p. 408. DMK-Press, Moscow (2017)
8. Esmaeili Gookeh, M., Tarokh, M.J. *J. Ind. Eng. Manage. Stud. Summer Autumn*, 24 (2), pp. 85-102.Working paper (2017)
9. Tkachenko, A., Lavrentev, D., Denisenko, M., Kuznetsova, V. SHS Web of Conferences 141, 05001 (2022) MTDE 2022 https://doi.org/10.1051/shsconf/202214105001
10. Shih, Y.Y., Liu, C.Y. (2003) *J. Database Mark Cust. Strategy Manage.*, 11 (2), pp. 159-172.
11. Abrukov, V.S., Karlovich, E.V., Abrukov, S.V. *2010 Research Bulletin of the Australian Institute of High Energetic Materials*, 2, pp. 129-144. (2011)
12. Bondarenko, G.G., Andreev, V.V., Stolyarov, A.A., Tkachenko, A.L. Vacuum, 2002, 67(3-4), pp. 617–621. https://doi.org/10.1016/S0042-207X(02)00262-2
13. Abrukov, V.S., Troeshestova, D.A., Abrukov, S.V., Karlovich, E.V., Polykarpov, A.I. Conference Paper (electronic format): Tenth Int. Symp. on Special Topics in Chemical Propulsion (10-ISICP) *At Ensma-Poitiers (France)*, p. 21. (2014)
14. Prokopenko, O., Larina, Y., Chetveryk, O., Kravtsov, S., Rozhko, N., Lorvi, I. *International Journal of Innovative Technology and Exploring Engineering*, 8 (12), pp. 4982-4987. (2019) doi: 10.35940/ijitee.L3745.1081219
15. Cuesta, H., Kumar, S. *Practical Data Analysis*. Packt Publishing Ltd, Birmingham (2016)
16. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. EMC Education Services. Wiley, Indianapolis (2015)
17. Gorbatiuk, K., Mantalyuk, O., Proskurovych, O., Valkov, O. *Proceedings of the 6Th International Conference on Strategies, Models and Technologies of Economic Systems Management (SMTESM 2019), Pp. 271–276 (2019)*
18. Gruzdev, A.: Prognoznoe modelirovanie v IBM SPSS Statistics i R: metod derev'ev reshenij [Predictive Modeling in IBM SPSS Statistics, R and Python: decision trees and random forest method], 274 p. DMK Press, Moscow (2018).

19. Katsko, I.A., Paklin, N.B.Praktikum po analizu dannyh na komp'yutere [Workshop on Data Analysis on a Computer: Training manual for universities] *276 P. Koloss* Gorelova, G.V. (ed.), Moscow, In Russian (2009)

20. Ткаченко, А. Л. Применение систем управления проектами при построении модели проекта / А. Л. Ткаченко, Р. А. Испирян // Математическое моделирование в экономике, управлении и образовании : сборник научных статей по материалам III Международной научно-практической конференции, Калуга, 16–17 ноября 2017 года. – Калуга: ООО "ТРП", 2017. – С. 86-92. – EDN YLXZLM.

21. Павлюк, А. Я. Системы электронного документооборота и управление отношениями с клиентами / А. Я. Павлюк, А. Л. Ткаченко // Актуальные вопросы современной науки : сборник статей по материалам XVIII международной научно-практической конференции, Томск, 13 февраля 2019 года. Том Часть 1(2). – Томск: Общество с ограниченной ответственностью Дендра, 2019. – С. 95-99. – EDN ZBPUSL.