UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Customer Lifetime Value for Retail Based on Transactional and Loyalty Card Data

Author:
Anastasiia KASPROVA

Supervisor:
Liubomyr BREGMAN

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Anastasiia KASPROVA, declare that this thesis titled, "Customer Lifetime Value for Retail Based on Transactional and Loyalty Card Data" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Fit fabricando faber."

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Customer Lifetime Value for Retail Based on Transactional and Loyalty Card Data

by Anastasiia KASPROVA

Abstract

The modeling of CLV in retail is a complicated task due to the lack of access to historical data of purchases, the difficulty of customer identification, and building the historical reference with a particular customer.

In this research, historical transactional data were taken from twelve North American brick-and-mortar grocery stores to compare different approaches to CLV modeling in terms of segmentation and forecast. Data engineering pipeline was applied to raw transactional data to transfer it into ready-for-modeling datasets, providing with the logic of each obtained feature. K-Means, Gaussian Mixture Model (GMM), DBSCAN clustering algorithms were applied to customer segmentation. The best outputs of clustering samples were later tested in CLV modeling. Unexpectedly, the K-Means algorithm results overperformed both GMM and DBSCAN ones.

For CLV modeling, two main models were considered: Markov Chain probabilistic approach of changing purchase behavior over time alongside with econometric Time Series revenue forecast and Survival Analytics lifespan estimates. The suggestions on CLV estimation for the offline retail business case were derived after result comparison with given advantages and limitations of each approach. Markov Chain model was suggested to check the general picture of the ongoing processes from the long-term perspective. On the other hand, Time Series revenue forecasting with Survival Analytics lifespan estimates could be used to check the expectations for the nearest feature. Moreover, the business value of CLV estimates and its applications were shown on examples derived from the results of both models: defined the promising clusters, checked their stability, and how they were formed, what customers were at risk to churn.

Acknowledgements

I want to express my special thanks of gratitude to Liubomyr Bregman for the supervision, all the support throughout the research, and hours spent on knowledge sharing and valuable feedbacks. I want to thank Dima Fishman, Sergii Shelpuk, and Oleksandr Romanko. I do appreciate your impact on my professional development choice. My special thanks go to Oleksii Molchanovskyi for the excellent quality Master Program in Data Science he has created at the Ukrainian Catholic University. I feel so lucky to know you in person and do value your humanity and professional qualities. I am incredibly thankful to Dmitry Leader and Grammarly Inc. for providing me with a scholarship to partly cover my tuition fees and reach the Dream much comfortable. And of course, I express my deepest gratitude to my parents who always support me with any idea or aspiration I have and to my son who always inspires me to Dream Big and Live in Wonders.

Contents

List of Figures

List of Tables

List of Abbreviations

CLV	**C**ustomer **L**ifetime **V**alue
CE	**C**ustomer **E**quity
CRM	**C**ustomer **R**elationship **M**anagement
NDA	**N**on-**D**isclosure **A**greement
ID	**I**dentification **N**umber
UPC	**U**niversal **P**roduct **C**ode
KNN	**K**-**N**earest **N**eighbors
DBSCAN	**D**ensity-**B**ased **S**patal **C**lustering of **A**pplications with **N**oise
EM	**E**xpectation **M**inimization
GMM	**G**aussian **M**ixture **M**odel
MM	**M**ixture **M**odel
ARIMA	**A**uto **R**egressive **I**ntegrated **M**oving **A**verage
SARIMA	**S**easonal **A**uto **R**egressive **I**ntegrated **M**oving **A**verage
ARIMAX	**A**uto **R**egressive **I**ntegrated **M**oving **A**verage with e**X**ogenous variables
SARIMAX	**S**easonal **A**uto **R**egressive **I**ntegrated **M**oving **A**verage with e**X**ogenous variables
VAR	**V**ector **A**uto**R**egressive
VARMAX	**V**ector **A**utoregression **M**oving **A**verage with e**X**ogenous variables
LSTM	**L**ong **S**hort-**T**erm **M**emory
CPH	**C**ox's **P**roportional **H**azard
RFM	**R**ecency **F**requency **M**onetary value
RNN	**R**eccurent **N**eural **N**etwork
ES-RNN	**E**xponential **S**moothing-**R**ecurrent **N**eural **N**etworks
GPU	**G**raphics **P**roccesing **U**nit
WCSS	**W**ithin **C**luster **S**um of **S**quares
RMSE	**R**oot **M**ean **S**um of **S**quares
AIC	**A**kaike **I**nformation **C**riterion
ADF	**A**ugmented **D**ickey–**F**uller test
ACF	**A**uto**C**orrelation **F**unction
PACF	**P**artial **A**uto**C**orrelation **F**unction
SVM	**S**upport **V**ector **M**achines
CART	**C**lassification **A**nd **R**egression **T**rees
NBD	**N**egative **B**inomial **D**istribution
MARS	**M**ultivariate **A**daptive **R**egression **S**plines
GAM	**G**eneralized **A**dditive **M**odel

Dedicated to my lovely family

Chapter 1

Introduction

1.1 Importance of CLV

CLV estimation helps management of a retail company with making data-driven decisions in the following areas (Villanueva and Hanssens, 2007, Jasek et al., 2018, Jasek et al., 2019): resources allocation and marketing strategies formulation, customer segmentation to build long-term relationships with clients, effective management of marketing investments, retaining and acquiring customers, estimation of company value and evaluation of marketing channels and campaigns.

According to Gupta et al., 2006, companies such as IBM, Capital One, LL Bean, ING, and others use CLV in managing and measuring their business success. The following factors cause their interest in CLV. (1) First of all, marketing metrics such as brand awareness, attitudes, sales, and share are not enough to represent the return on marketing investment. Marketing campaigns that improve sales or share can negatively impact a long-term profitability of a company. (2) Secondly, financial metrics such as stock price and aggregated profit of a company are restricted with a diagnostic capability. Aggregated financial metrics cannot catch unprofitable customers to allocate resources appropriately. (3) Development of IT enables companies to collect customer transactions easily, build and analyze models, convert data into insights, and customize marketing programs for individual customers.

1.2 Challenges of CLV modeling for offline retail business

CLV in offline retail business faces two significant challenges: customer identification to build the historical reference with a particular customer and a lack of publicly available transactional datasets to validate theoretical aspects of CLV modeling.

According to Jasek et al., 2018, Jasek et al., 2019, there is a number of papers covering theory of CLV modeling, but there are few of them dedicated to empirical analyses, and only a few (Batislam and Filiztekin, 2007, Nikkhahan, Habibi, and Tarokh, 2011, Jasek et al., 2018, and Jasek et al., 2019) used real transactional data in their research. Besides, blogs and tutorials (Medium, TowardsDataScience, etc.) or online courses (datacamp, coursera, edx, udemy) dedicated to customer analytics, mostly re-use the same datasets (for instance, Online Retail Dataset is often used for RFM analysis explanation, Iris Dataset – clustering, Telco – churn prediction, Titanic – survival analytics). The reason is that most research results in the retail field have a proprietary nature. Retailers rarely reveal results of their investigations to public access to stay on the top of the global market.

In the study, I have used new transactional and loyalty card data taken from the offline retail business. Having raw data, I have prepared different datasets according

to appropriate model requirements and demonstrated the most popular approaches of CLV modeling with it.

1.3 Goals of the master thesis

The practical goal of the project is to create an applicable analytical framework for offline and semi offline businesses to run CLV forecast for channels and campaign evaluations. The pre-requirement is to have transactional data of purchases, with a selection of suitable for offline retail environment CLV model application based on transactional and loyalty card data.

According to the goal, there are three main tasks to solve:

I. Investigate existing approaches to CLV modeling suitable for the offline retail business.

II. Develop a two-component CLV model: cluster customers and predict revenue relevant to the cluster of interest based on available transactional and loyalty card data.

III. Compare models based on their assumption limitations and advantages for an offline store environment.

1.4 Thesis Structure

Firstly, a classification of CLV estimation approaches is introduced in Chapter 2. Then, in Chapter 3, two approaches for CLV modeling and background information, that makes the basis of proposed methods, are considered. In Chapter 4, each component of the data engineering pipeline of dataset preparation is described. In Chapter 5 the environment where all the experimental work was conducted and the results of proposed models are represented and discussed. Finally, in Chapter 6, the suggestions on CLV estimation for the offline retail business case and a list of the directions of possible future work are represented.

Chapter 2

Overview of existing approaches

According to McKinsey, *Customer Lifecycle Management*, Customer Lifetime Value (CLV) analysis provides a 360-degree view of customers and is used as a tool for a value maximization on customer level increase in the revenue of a company. Kotler, 1974 defined CLV as "present value of the future profit stream expected given a time horizon of transacting with the customer"; however, later, according to Estrella-Ramón et al., 2013, it has studied under different names: Lifetime Value, Customer Equity, Net Present Value, Customer Profitability, and Customer Value. Various researches define CLV slightly differently, but the general idea is that this metric indicates the total revenue a company can expect from a single customer to generate over time. Once it is estimated, it is possible to identify promising customers, minimize acquisition cost by specific targeting of customers, and organize an efficient CRM (adjust promotions, recommendations, customer service) by prioritizing customers (*Machine Learning for Marketing Analytics in R* course notes).

2.1 CLV Models Classification

According to Gupta et al., 2006, as well as later mentioned by Estrella-Ramón et al., 2013, Jasek et al., 2018 and Jasek et al., 2019, there are two approaches to examine CLV components: deterministic and stochastic, combining six different methods of modeling together.

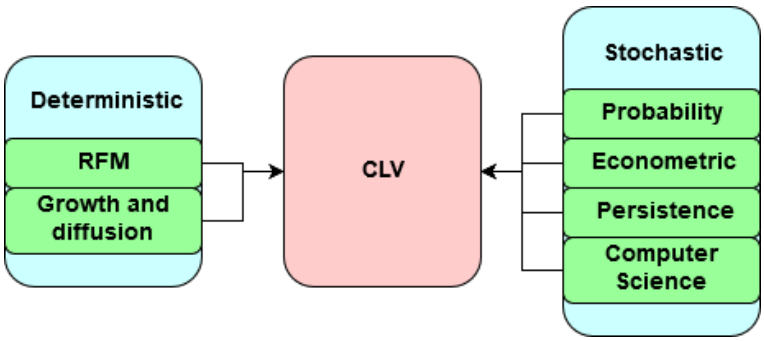


FIGURE 2.1: CLV models classification defined by Gupta et al., 2006.
Source: Author

2.1.1 Deterministic Approach

The deterministic approach uses equations where all the metrics are entered directly in the simplified calculation of CLV. The formulas are easy to use for managers; however, the model is purely descriptive and ignores the heterogeneity of individual

customer response probabilities (retention, churn rate within a cohort). The representatives of this approach are RFM (Recency, Frequency, Monetary value) and Growth and Diffusion models.

RFM

RFM models have been popular in direct marketing for more than 30 years. These models were developed to improve response rates by targeting marketing programs at specific customers. Before that, companies used the demography of customers. These models describe customer behavior based on three variables: recency – time since the last transaction, frequency – number of transactions during a given period, monetary value – the total amount of money spent in transactions during a given period (Hughes, 2011). The simplest models segment customers based on each value for these three variables. More sophisticated models use weights making RFM variables more or less important.

Despite the simplicity of RFM models, they have multiple limitations: the customer behavior can be predicted only for the next period, a precise description of the actual customer behavior cannot be derived from the model variables, other variables are not taken into account, a monetary amount of customer value is not offered as a model output. To deal with the disadvantages of the deterministic approach, some researchers advise to mix this approach with stochastic one (RFM with Markov Chain, for instance, Pfeifer and Carraway, 2000). In the study, RFM variables are used as input to other models.

2.1.2 Stochastic Approach

The stochastic approach characterizes a sequence of random variables which are integrated into another variable. Each random variable has its probability distribution function. The variables can correlate. These CLV models are considered more accurate since they take into account customer heterogeneity (retention and churn rate). According to Gupta et al., 2006, Estrella-Ramón et al., 2013, Jasek et al., 2018 and Jasek et al., 2019, all stochastic approaches can be divided into probability, econometric, persistence, and computer science models.

Probability models

The probability model is a representation of the world where observed behavior is modeled as a stochastic process specified by unobserved/latent behavior, which differs among customers according to some probability distribution. Models that belong to this category are Pareto/NBD (Negative Binomial Distribution), Hierarchical Bayesian, and Markov Chain (Gupta et al., 2006, Estrella-Ramón et al., 2013). In the study, a Markov Chain model to determine the path of a particular customer of switching the clusters over time and then calculate its CLV has used.

Econometric models

Econometric models are similar to probabilistic ones and are widely used in the industry due to their simplicity. These models consist of submodels corresponding to customer acquisition, retention, and expansion (cross-selling and margin), and their output later used in CLV estimations. In this work, the focus was on finding a retention component (a probability that a customer stays with a company and performs purchasing again). Thus, it has been identified if a customer is still "active"

(make purchases from a company) and predicted their lifetime duration with a company. Such survival models as Kaplan-Meier survival curves and Cox's proportional hazard have been used to estimate time to event of a customer churn from the relationship with a company, as well as time series models to estimate the revenue a particular customer or cluster can bring to the store.

Computer Science models

To computer science models, Gupta et al., 2006 include algorithms of data mining, machine learning, and statistical non-parametric statistics, which use a large number of variables and have high predictive ability. These are projection-pursuit, neural network, decision tree models, spline-based such as generalized additive models (GAM), multivariate adaptive regression splines (MARS), classification and regression trees (CART) and support vector machines (SVM), Random Forest, Extreme Gradient Boosting and other classical Machine Learning Approaches. However, he states that many of them could be rather suitable for customer churn studies, than CLV modeling (referring to his shared work with Neslin and Kamakura (Neslin et al., 2006), and points out that these approaches should be studied in future.

Many of the models, developed by scientists decades ago, usually have a different context for different types of business as well as their investigation was based on various data sources (database of customers, surveys, public reports, panel data, and managerial judgments, according to Estrella-Ramón et al., 2013). In works of Jasek et al., 2018 and Jasek et al., 2019 it is stated that probability and econometric approaches are only suitable for retail business. Thus, for their purposes, they select an extended Pareto/NBD model with RFM factors in the computations and Markov chain model. They do not include diffusion/growth and persistence model to their research, because these models are not applicable at the level of individual customers prediction, but used more in Customer Equity (a long-term value of a company which is calculating as a sum of CLV of all customers). The authors do not deal with computer science models as well, mentioning that these models are not enough considered in the literature with a focus on CLV calculations.

2.2 Customer segmentation

In practice, the revenue estimation is rarely done on the individual level due to a low level of accuracy. Instead, data analytics/data scientists usually divide customers by their similar purchase behavior into groups (micro-segmentation) and then predict future purchases towards a particular segment of customers.

There are multiple both unsupervised and supervised approaches to a customer segmentation described by Wedel and Kamakura, 1999. All of the methods which belonged to unsupervised machine learning have no exact answer on how to divide customers, and the results of segmentation could vary from approach to approach. In the study, the most common clustering algorithms have been used, particularly K-Means, Density-Based Spatial Clustering (DBSCAN), and Gaussian mixture models (GMM).

Chapter 3

Proposed models

The main aim of this study is to build a CLV model based on transactional and loyalty card data. To achieve this goal, different approaches have been examined and the one, which fits the data best, has been selected:

1. Probabilistic approach with Markov Chain revenue estimation (Fig. 3.1).

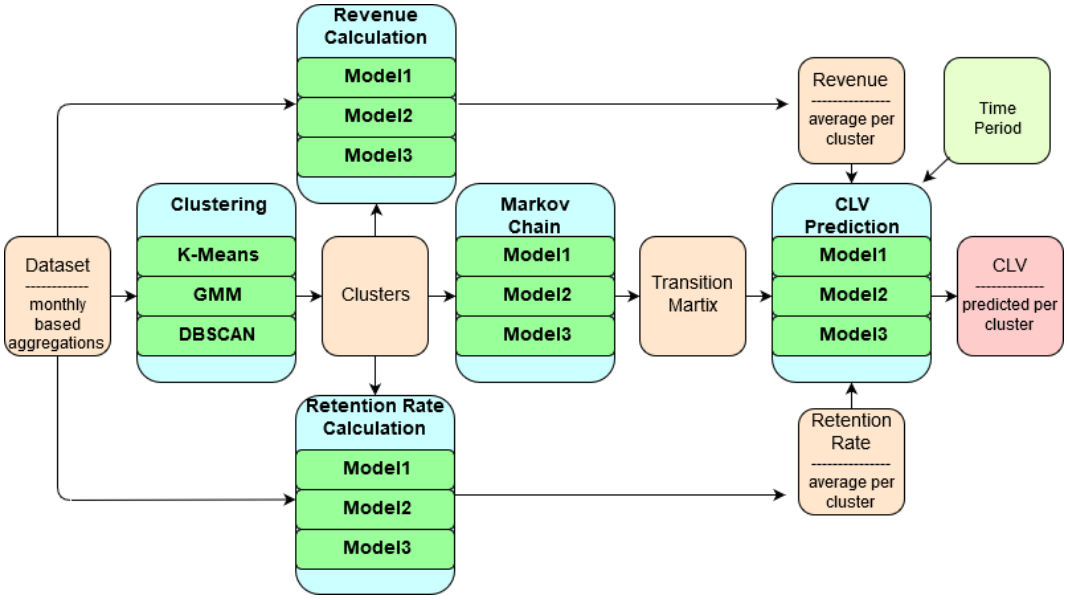


FIGURE 3.1: Pipeline of a probabilistic approach with Markov Chain CLV prediction. Source: Author.

In this approach, I prepared a dataset based on 4-weeks analytics of each customer (see Chapter 4, Dataset Preparation) and clustered customers using K-Means, Gaussian Mixture Model, and DBSCAN clustering algorithms. Then transition matrices has been calculated using Markov Chain Model for each of the clustering outputs. Knowing IDs of households of each cluster, I derived average revenue and average retention per cluster. After that I predicted CLV for each cluster for the period I have been interested in and evaluated those three models on actual historical data.

2. Econometric approach with survival analytics lifetime estimation and time series revenue prediction (Fig. 3.2).

In this approach, I prepared two datasets separately with weekly aggregations on a customer level for Time Series Forecasting and overall period aggregations for Survival Analytics. To predict revenue per week on an individual level, I chose the best Time Series Model from ARIMA/SARIMA and LSTM based on AIC metric. I checked the overall customer life duration with a Kaplan-Meier survival curve and

estimate individual lifespan with Cox' Proportional Hazard Model for each household with a 0.5 threshold. Finally, the outputs of both models have been used to calculate CLV for each household.

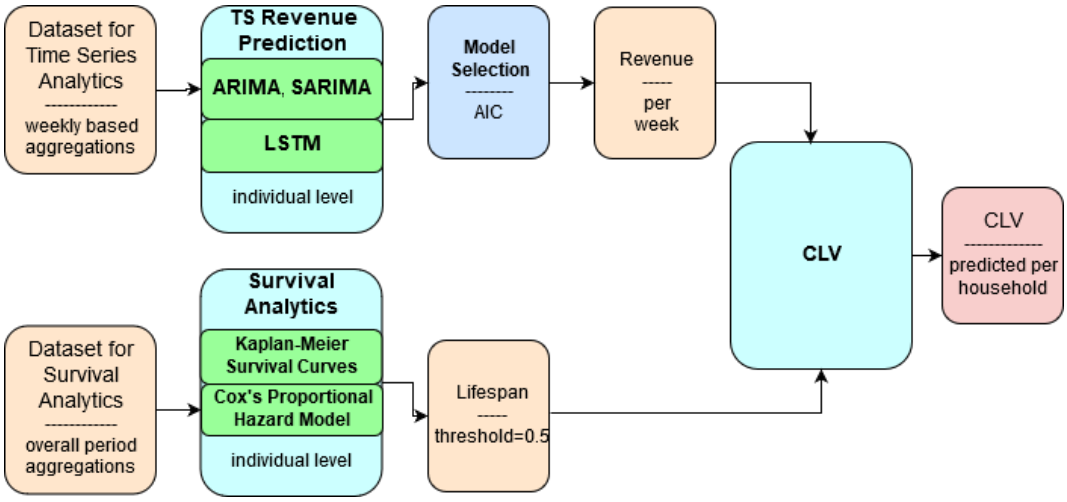


FIGURE 3.2: Pipeline of the econometric approach of CLV prediction.
Source: Author.

Once the results of both approaches ware obtained, it became possible to conclude what method with what data pipeline works best for the data used in the research; also, there was one more step added to the econometric approach – mapping cluster IDs to households and month IDs to weeks.

3.1 Clustering Algorithms

Clustering is a process of applying unsupervised Machine Learning algorithms for automatic grouping of objects where objects within each group (cluster) are more similar to each other than objects from different groups. By belongings of data points to a cluster, there are two subgroups of clustering that can be considered: hard, when each data point belongs only to one cluster, and soft, when each data point is an assigned probability to be in clusters. By approaches, clustering algorithms can be divided into four types: centroid, distribution, density, and connectivity models. Bellow, I consider the representatives of the first three groups to check what approach works best with the available dataset.

3.1.1 K-Means

K-Means is a centroid-based clustering algorithm, which is based on distance measurement between data points and a centroid of the cluster.

The objective function is defined as:

$$arg\,min \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu||^2 \tag{3.1}$$

where x – data point of the dataset, μ – cluster mean, S – cluster, k – number of clusters.

To partition data points into k predefined clusters, K-Means initializes arbitrary k data points – centroids (cluster centers) and assigns each data point to the closest

centroid forming k cluster, measured by Euclidean distance, cosine similarity, etc. Then each cluster center is replaced by the coordinate-wise average of all data points that are the closest to it (Fig. 3.3). Both steps are repeated until convergence. K-Means algorithm converges to a local minimum of the within-cluster sum of squares (Lloyd, 1982).

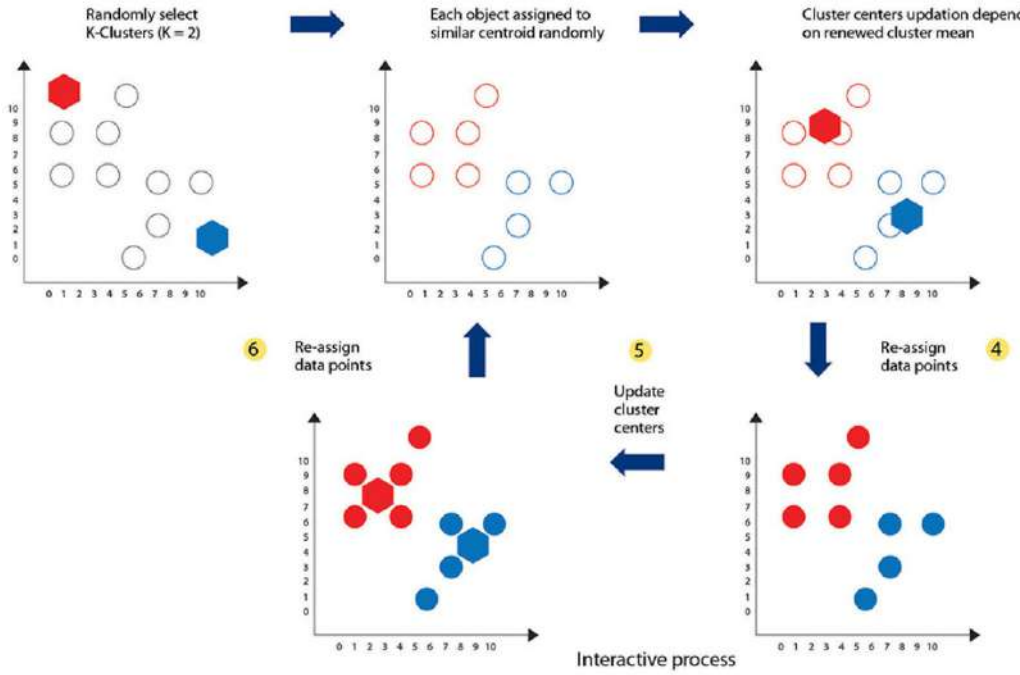


FIGURE 3.3: Illustration of K-Means clustering steps. Source: *K-Means and X-Means Clustering*.

The K-Means algorithm is popular due to its computational efficiency and ability to handle large datasets. However, it provides only linear separation of cluster boundaries, expects that the number of clusters is known, does not work with categorical, ordinal or count variables, and not very robust to missing data. It can also be hard to maintain over time and replicate the results.

3.1.2 Gaussian Mixture Model

Gaussian mixture model (GMM) is a probabilistic approach of clustering. It uses a soft clustering approach for distributing the points in different clusters.

According to probabilistic interpretation, clustering can be considered as the identification of components of a probability density function that generated the data. Similarly, an identification of cluster centroids can be considered as a discovery of the modes of distribution. Thus, GMM assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters (Fig. 3.4). Mixture Models can be also considered as a generalization of K-Means clustering to incorporate the information about the covariance structure of data as well as centers of latent Gaussians.

Mathematically a mixture of k Gaussians can be represented by a formula:

$$p(x) = \sum_{i=1}^{k} \pi_k N(x|\mu_k, \Sigma_k) \tag{3.2}$$

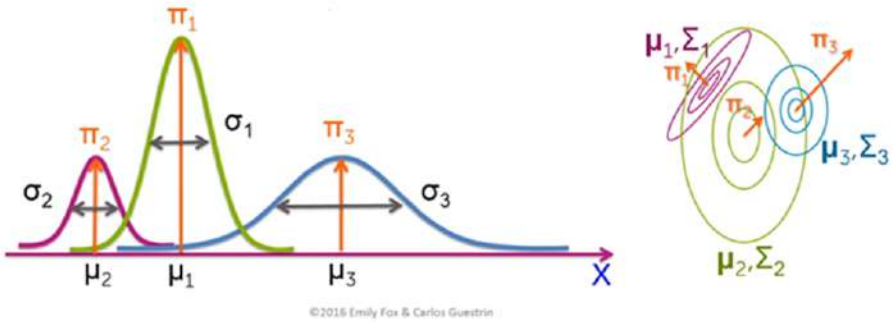


FIGURE 3.4: Gaussian Mixture Model representation.
Source: Fox and Guestrin, 2016.

where $N(x|\mu_k, \Sigma_k)$ - cluster with centroid μ_k and variance Σ_k (mixture components), π_k - weights or mixture proportion.

To find the parameters of a probabilistic model such as GMM, the Expectation-Maximization (EM) algorithm is commonly used (Bishop, 2006). Qualitatively EM does the following: it chooses to start guesses for the location and shape and then iterates until convergence $E - step$ (for each point, find weights encoding the probability of membership in each cluster) and $M - step$ (for each cluster, update its location, normalization, and shape based on all data points, making use of the weights).

The GMM approach of clustering is flexible in terms of cluster covariance, and, according to the method, data points can belong to different clusters simultaneously. However, it does not work well with high dimensional data, as it can miss the globally optimal solution (as K-Means), and requires the number of MM (clusters) to be specified.

3.1.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering non-parametric algorithm. DBSCAN algorithm groups together data points that are closely located (data points with many nearest neighbors), making as outliers data points that are located in the low-density regions (data points whose nearest neighbors are too far away).

To detect a cluster, DBSCAN starts with an arbitrary data point p and retrieves all data points density-reachable from the data point p with respect to two parameters: ε (a distance which specifies how close data points should be to each other to be considered as a part of a cluster) and $minPts$ (minimum data points required to form a cluster). Using these two parameters, DBSCAN splits all data points into three categories: core points, border points, and outliers (Fig. 3.5). If p is a core point, then it forms a cluster with all points (core or non-core) that are reachable from it. If p is a border point, then no points are considered as density-reachable from p, and DBSCAN proceeds with another point of the dataset (Ester et al., 1996).

In the diagram represented on Fig. 3.5, $minPts = 4$. Point A and the other red points are core points because the area surrounding these points in a ε radius contains of at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable(Schubert et al., 2017).

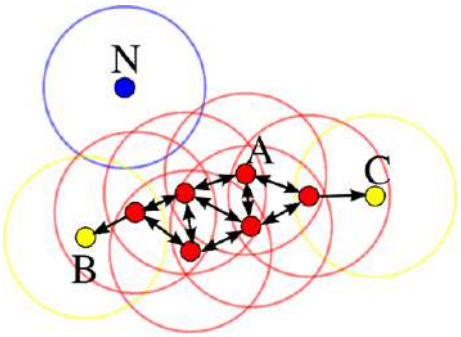


FIGURE 3.5: DBSCAN slit explanation. Source: Schubert et al., 2017.

In contrast to K-Means and GMM algorithms, DBSCAN discovers the number of clusters. It can also detect the outliers, supports a non-linear separation of the clusters and can separate high- and low-density clusters. However, DBSCAN can be completely inappropriate for certain datasets, because it tends to merge clusters with overlapping regions. Moreover, it still requires to set up two parameters: ε and *minPts*.

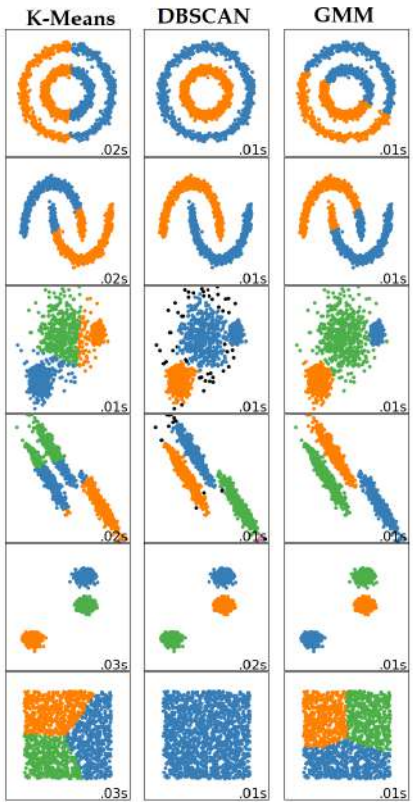


FIGURE 3.6: Comparison of K-Means, DBSCAN and GMM clustering algorithms on 2D data. Source: *Overview of clustering methods*.

The comparison of the clustering algorithms performance used in the study is represented in Fig. 3.6. However, data used for algorithm performance testing had only two dimensions.

3.1.4 Summary

There is no such thing as the best clusterization technique. It is considered that the best one as the one, which is the most appropriate to the business requirements and the data. Most clusterization algorithms are sensitive to the input dataset. They require relatively normally distributed standardized variables as an input. Moreover, the performance of any clustering algorithm is heavily influenced by defined parameters. In each case there are multiple techniques to help with their choice (Elbow Method, Silhouette score, Gap Statistics, etc.).

3.2 Markov Chain

Markov chain is a probabilistic model that describes a sequence of possible events where the probability of transition from one state to another depends only on the state of the previous event (Pfeifer and Carraway, 2000). The probability of transition can be considered as a transition matrix $T[i, j]$, which represents the percentage of customers who moved from segment i to segment j within some period.

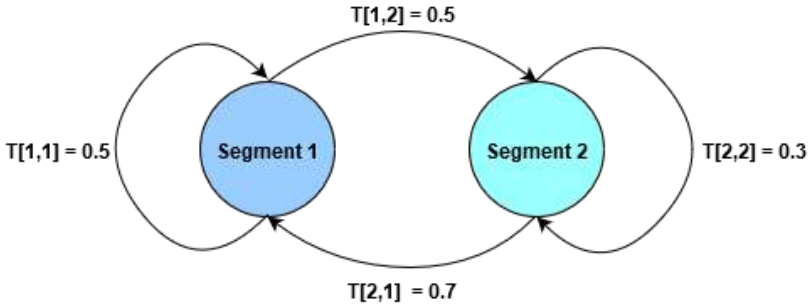


FIGURE 3.7: Example of the transition graph between segment 1 and segment 2. Source: Author.

According to the transition graph (Fig. 3.7), the corresponding transition matrix can be represented as

$$T = \begin{bmatrix} 0.5 & 0.5 \\ 0.7 & 0.3 \end{bmatrix} \tag{3.3}$$

The estimation of the probability of transitions during the future period of the time equals T^n, where n is a number of sequential units of time (months/years). Thus, knowing the transition matrix, one can estimate the probability of the customer's future moves from one segment to another.

Markov Chain method belongs to analytical methods, where parameters of the system are calculated by a particular formula, which makes it easy to apply. However, it only considers the latest state of the system, does not take into account the history of previous states, and remains constant over time (Haenlein, Kaplan, and Beeser, 2007).

3.3 Survival Analysis

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen (Miller, 2011). In the retail domain, the considered event corresponds to the customer churn (when a customer stops any transactional

relationships with a company), and the duration of time is a lifetime of customer's relationship with a company (from the first transaction to the last one in the store).

Survival analysis is used in several ways: to describe the survival time of an individual or a group, to compare the survival times of two or more groups, or to describe the effect of categorical or quantitative variables on survival (Miller, 2011). In the study, I am interested in the estimation of customer's lifespan that can be measured in days, months, years, depending on model configurations. The key benefit of using survival analysis for lifetime estimations is that it deals with censored data (when a subject has no events during the time of the observation).

3.3.1 Kaplan-Meier Survival Curves

Kaplan-Meier survival curves (Kaplan and Meier, 1958) (Fig. 3.8) are widely used in the estimation of survival function $\hat{S}(t)$, probability that an individual survives at the end of a time interval, on the condition that the individual was present at the start of the time interval.

$$\hat{S}(t) = \prod_{t_i t}^{n} \frac{n_i - d_i}{n_i} \tag{3.4}$$

where $\hat{S}(t)$ - a survival function, t_i – duration of study at point i, n_i – number of individuals at risk just before t_i, d_i – number of customers who did not survive.

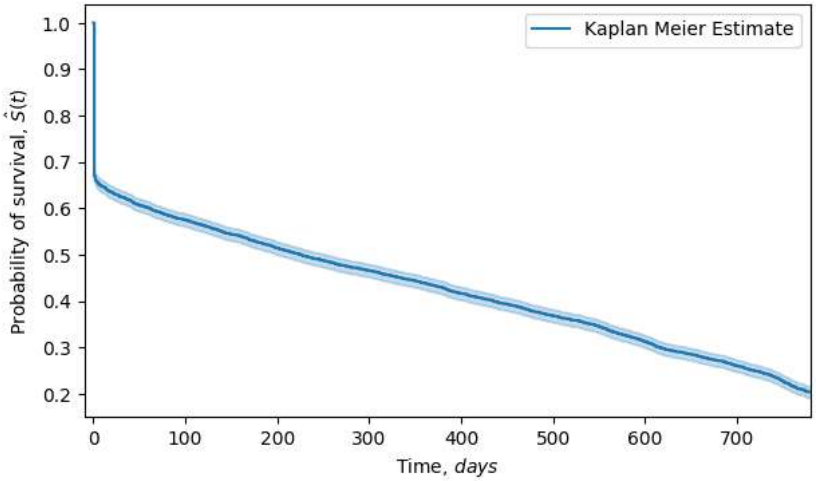


FIGURE 3.8: Survival curve estimated by Kaplan-Meier model. Source: Author.

Kaplan-Meier method requires a minimal feature set to build survival curves: the time when an event occurred and the lifetime duration between birth and an event. Moreover, it handles class imbalance automatically. Since it is a non-parametric method, few assumptions are made about the underlying distribution of the data; however, it can not be as efficient or accurate as alternative methods on problems where the underlying data distribution is known. Also, it assumes the independence between censoring and survival, thus at time t is both censored and non-censored have the same estimations.

In the study, Kaplan-Meier survival curves were used to overview the expected lifetime duration of customers.

3.3.2 Cox's Proportional Hazard Model

Cox's Proportional Hazard Model is a regression model (Cox, 1972) used for investigating the effect of several variables upon the time a specified event takes to happen. Its central assumption that the impact of the predictor variables upon survival is constant over time and additive on one scale.

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \ldots + \beta_p X_{ip}) \tag{3.5}$$

where $\lambda_0(t)$ – baseline hazard function, describing how the risk of event per time unit changes over time as baseline levels of variables, $\exp(\beta_1 X_{i1} + \ldots + \beta_p X_{ip})$ - partial hazard function (effect parameters), describing how the hazard changes in response to the explanatory covariate. Since Cox's Proportional Hazard model has the time component in the baseline hazard function, variables can only affect the risk prediction by increasing or decreasing the baseline value.

Although using Cox's Proportional Hazard model, it is possible to build survival curves for each subject with respect to all used variables, all survival curves representing different subjects have the same basic shape.

In this work, Cox's Proportional Hazard model was used to estimate the customer probability to churn in a particular time t on the individual and cluster level of aggregation.

3.4 Time Series

Time Series is an ordered sequence of values of a variable at equally spaced time intervals (irregular data does not form time series). Time Series Analysis provides a variety of methods for analyzing time series data to extract meaningful components – trend, seasonality, cyclicity, and irregularity. Time series forecasting is the use of a model to predict future values based on the previous ones. For the research, ARIMA/SARIMA and LSTM forecasting methods were selected due to their popularity.

3.4.1 ARIMA, SARIMA

Auto Regressive Integrated Moving Average model (ARIMA(p,d,q)) is one of the most widely used time series forecasting approaches.

Autoregressive model forecasts correspond to a linear combination of previous values of the variable:

$$AR(p) : y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + \varepsilon_t \tag{3.6}$$

Moving Average model forecasts correspond to a linear combination of previous forecast errors.

$$MA(q) : y_t = m_1 \varepsilon_{t-1} + m_2 \varepsilon_{t-2} + \ldots + m_q \varepsilon_{t-q} + \varepsilon_t \tag{3.7}$$

Since both models require time series data to be stationary, the Integrated part (differencing) takes care of it:

$$I(d) : \Delta y_t = y_t - y_{t-1} \tag{3.8}$$

ARIMA model can be extended to Seasonal Auto Regressive and Seasonal Moving Average parts SARIMA(p,d,q)(P,D,Q)s, where p – autoregressive order, d – differencing order, q – moving average order, P – seasonal autoregressive order, D – seasonal differencing order (subtracting time series value of one season ago), Q – seasonal moving average order, S – number of steps per cycle (*Forecasting using ARIMA models in Python* course notes). For example,

$$ARIMA(2,0,1) : y_t = a_1 y_{t-1} + a_2 y_{t-2} + m_1 \varepsilon_{t-1} + \varepsilon_t \tag{3.9}$$

$$SARIMA(0,0,0)(2,0,1)_7 : y_t = a_7 y_{t-7} + a_{14} y_{t-14} + m_7 \varepsilon_{t-7} + \varepsilon_t \tag{3.10}$$

ARIMA and SARIMA models require only a target variable (however, there are extended version of these models with exogenous variables: ARIMAX and SARIMAX). If a model fit well to the data and overall trend doesn't change, it can successfully forecast small ups and downs; however, if an unusual growth or fall is observed, ARIMA fails with its forecasting (Brownlee, 2018).

3.4.2 LSTM

Long short-term memory (LSTM) is a modified version of a recurrent neural network (RNN) that can be applied for time series forecasting. LSTM trains the model with backpropagation, vanishing gradient problem is solved, each repeating module has four layers (instead of one as a standard RNN has).

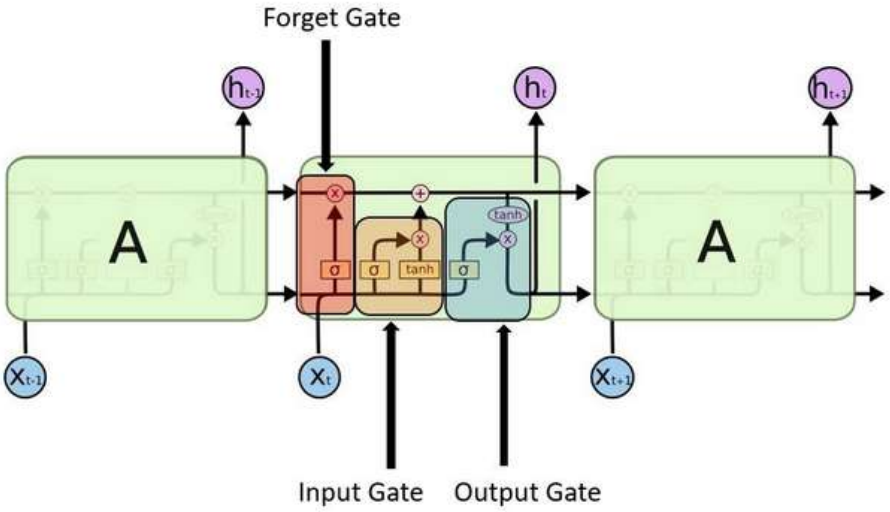


FIGURE 3.9: Repeating module in LSTM. Source: Mittal, 2019.

According to Fig. 3.9 architecture of LSTM, there are three gates present in each module of the network (Mittal, 2019). Forget gate – discover what details should be discarded from the module. It considers the previous state h_{t-1} and X_t, as input data and outputs the number between 0 and 1 for each number in the cell state C_{t-1}.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f \tag{3.11}$$

Input gate – discovers what value from the input data should be used to modify the memory.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \tag{3.12}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C \tag{3.13}$$

Output gate – considers the input and block memory for providing with an output.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o \tag{3.14}$$

$$h_t = o_t * tanh(C_t) \tag{3.15}$$

LSTM is a nonlinear model that learns nonlinearity from the data. It might require a sample of a particular size to learn non-linearity and longer training time compared to linear models such as ARIMA/SARIMA. However, its predictions are considered to be more accurate.

3.5 CLV calculation

CLV is a total revenue a company can expect from a single customer or a cluster of customers to generate over time. It consists of $CLV_{current}$ - the revenue a customer has already brought to the company (extracted from historical data) and CLV_{future}– the revenue expected to be brought to the company (estimated with modeling).

$$CLV = CLV_{current} + CLV_{future} \tag{3.16}$$

Depending on objectives, CLV_{future} can be considered as a forecast not for all the period of the relationship between a company and a customer but for the next few months. The CLV formula can vary for different business domains, available data and approaches of modeling. Since in the transactional dataset the information about profit or margin is excluded, the following formulae of CLV_{future} can be used for estimations:

Markov Chain

$$CLV_{future} = \sum_{n=1}^{N} T^n R_n r_n \frac{1}{(1+d)^n} \tag{3.17}$$

Time Series and Survival Analytics

$$CLV_{future} = \sum_{n=1}^{N} R_n P_n \frac{1}{(1+d)^n} \tag{3.18}$$

where d – discount rate, N - period of interest, T - transition matrix, r - retention rate, P - probability to survive, R - revenue, $\hat{R}$ - revenue estimate.

Discount rate can be taken as the Cost of Capital for grocery and food retail business as 4.35% per year (Damodaran, 2020). However, for roughly estimates of CLV in retail it is often omitted.

3.6 Summary

In this chapter, two different approaches with the corresponding pipelines were proposed for the CLV modeling based on transactional and loyalty card data: a probabilistic Markov Chain and an econometric Survival Analytics together with a Time Series forecasting. Moreover, the background information according to each method used in the pipelines with its advantages and disadvantages was provided.

Chapter 4

Datasets Preparation

For CLV study the raw transactional and loyalty card data was obtained from the technical department of the grocery store. To be able to use it for CLV modeling the data was passed thought the dataset preparation pipeline (Fig. 4.1), which consisted of four phases: data cleaning, feature generation, dimentionality reduction, and data preprocessing (according to the requirements of the algorithms used for modeling).

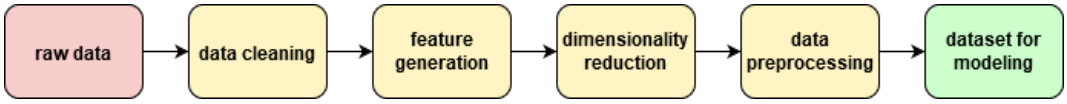


FIGURE 4.1: General dataset preparation pipeline. Source: Author.

4.1 Raw Data Description

Raw transactional data and loyalty card data of twelve stores of a large grocery retail chain was obtained from the industry partner. Due to the signed NDA, I can disclose only that the location of the stores is in North America. These stores offer a variety of grocery products and provide their customers with loyalty cards, linked to particular households. Each store issues product discounts and sends them to cardholders by regular mail. The loyalty program seems to be popular since about 85% of store revenue and about 80% of transactions are obtained from the cardholders.

The raw transactional data has 25 months history of purchases (from September 2015 to October 2017) of more than 500 thousand unique customers and includes the following information: a transaction ID, a date of purchase, a UPC of an item, quantity, a regular price, a discounted sale price, and a loyalty card number (helps to track an individual customer behavior over time).

While exploratory data analyses I noticed the following challenges with the data: some households had more than one active card, multiple customers had only 1-2 purchases, not all UPCs were assigned to a category/subcategory of products, the group of goods produced by the store had internal classification on UPC-level and the granularity of classification was too detailed.

4.2 Data Cleaning

Different approaches to modeling CLV require different features and data preprocessing. However, data cleaning remains the same for each of the models. I decided to keep only purchases of cardholders, associating each purchase with a corresponding household, drop items in the transactions when a category was missed and manually group categories to a higher level of granularity.

4.3 Feature generation

According to selected for the study models, there were three different datasets prepared based on transactional and loyalty card data: for clustering (the results of which were later used in Markov Chain, Survival Analytics and Time Series) on 4-weeks aggregated level, for Survival Analytics with aggregation on lifetime duration level and Time Series on daily and weekly aggregation level (time series of monthly data was excluded from the research due to the size of the time period available).

4.3.1 Clustering

The choice of features for clustering was motivated by the idea that the attention should be paid not only to the revenue from an individual customer or a cohort but also to the potentially interesting for customers category of products.

• RFM: *recency* - how recently the household purchase, referring to the number of days which had passed since the last time a household purchased some goods from the store; $frequency$ - how often the household purchase, referring to the number of invoices with purchases during the particular month; $monetaryvalue$ - is the amount of money the household spends during a particular month.

• Churn: all the history of household' purchases was considered. If the last purchase was done in the previous to the considered month of purchase and it was not the last historically available month, then this household was 'churned'.

• Discounts: number and monetary value of store coupons, manufacturer coupons, refunds a household applied during the month.

• Categories of Products Purchased - the proportion of money a household spent on each category of products.

• Loyalty program features: duration since a household registered the first loyalty card; the number of loyalty cards used by a household.

4.3.2 Survival Analytics

Kaplan-Meier Survival Curves

• Churn: same logic as applied above.

• Duration: the period in days between the first and the last transactions.

Cox's Proportional Hazard Model

• Churn: same logic as applied above.

• Duration: the period in days between the first and the last transactions.

• Monetary Value: the amount of money the household spends during a relationship with a store.

• Frequency: the same logic as above.

• Mean receipt.

• Standard Deviation of a receipt.

• Days since the loyalty card was issued.

4.3.3 Time Series

ARIMA/SARIMA/LSTM

• ID: household ID or cluster ID.

• Date: day or beginning of the week.

• Monetary Value: the amount of money a household or a cluster spent in the store during a particular period – day or week.

4.4 Data Preprocessing

Some clustering algorithms (K-Means, GMM) have requirements for the input data: the variables should be from a symmetric distribution (not skewed) and have the same average values and variance.

To check the skewness of the data, the distribution of each feature was built. The features with the left skewness were transformed into the logarithmic scale to reshape the distribution into more or less symmetric.

To make the data have the same average values and variance, all features were transformed using a standard scaler:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

where x - value of a variable, μ – mean, σ – standard deviation. Otherwise, input variables with larger variances would affect the results of clustering.

4.5 Dimensionality Reduction

In order to avoid misusage of algorithms and biases I checked if the data needed dimensionality reduction, the matrix of correlation was built (Fig. 4.2). However, no strongly correlated features (with a correlation coefficient > 0.9) were detected that could possibly affect the clustering results. Moreover, since the dataset contained only 21 features, no dimensionality reduction technique was applied.

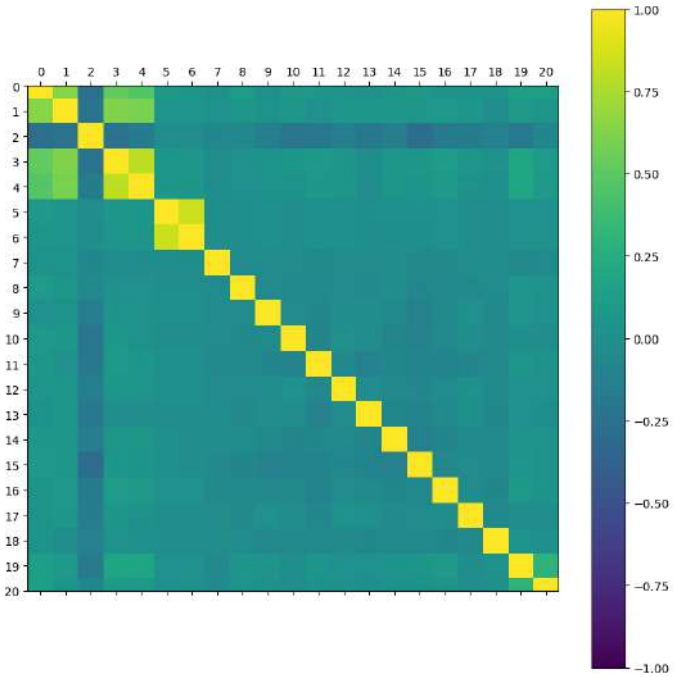


FIGURE 4.2: Feature Matrix Correlation. Source: Author.

4.6 Summary

The full data engineering cycle of the datasets preparation is considered in this chapter – from the raw data obtained from the company's technical department to

datasets used in different approaches of modelling CLV. Each step of the data preparation pipeline is provided with a detailed explanation. Moreover, the description of each feature logic in the datasets is provided.

Chapter 5

Experiments

5.1 Analytical Environment

The data analysis was executed in the PySpark environment on Google Colab, exploiting their free available GPU (K80 in most cases). I had to contribute some time and effort on rewriting all the scripts in PySpark syntax, but it helped me not only with speeding up the processes (mostly feature generation, feature selection, data preprocessing before clusterization), but also be on the same page with a scientific supervisor. Markov Chain, survival analysis and time series forecasting were done in Python. Markov Chain module was written by me, Time Series ARIMA forecasting was modeled by statsmodels package (Seabold, Skipper, and Perktold, 2010) and Survival Analysis was conducted with the help of the lifelines package.

Additionally, there was an attempt to launch NVIDIA's Rapid in Google Colab; however, I was unlucky to receive access to T4 GPU permanently.

5.2 Clusterization

The preliminary idea of the clustering application in the research is that I have detected different behavioural patterns and group all the customers into clusters based on their historical purchases (taken into account not only revenue and store loyalty but customers' basket preferences), investigate purchase activity of each cluster and predict its value for the store in the future (CLV). The exact data preparation for clusterization techniques is described in Section 4.3.1.

5.2.1 K-Means

To conduct customer clusterization by applying the K-Means algorithm, the number of centroids should be parametrized. To solve this puzzle, I applied K-Means clustering technique for range of possible number of clusters between 0 and 50, measuring Within Cluster Sum of Squares (WCSS) (Fig. 5.1), and a Silhouette Score (Fig. 5.2) as well.

Elbow Method

The algorithm scans through the estimated range of clusters and calculates the sum of squared distances within clusters (WCSS) - how far a point from a cluster it is assigned to. Then the dependence of WCSS on the number of clusters is plotted, and a value (appropriate cluster number) where the curve flattens out is indicated.

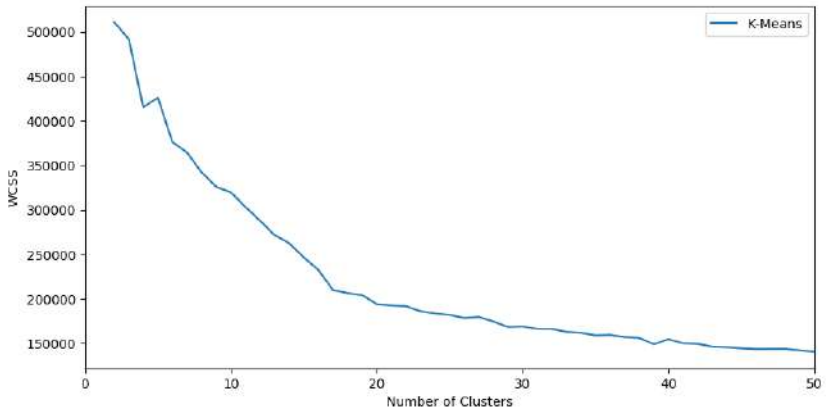


FIGURE 5.1: Dependence of Within Cluster Sum of Squares (or K-Means score) on the number of clusters (Elbow Method). Source: Author.

Silhouette Score

According to *Sklearn silhouette score documentation*, silhouette score (SS) is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample and evaluates how close each data point is to its cluster versus how close it is to the other clusters.

$$SS = \frac{b - a}{max(a, b)} \tag{5.1}$$

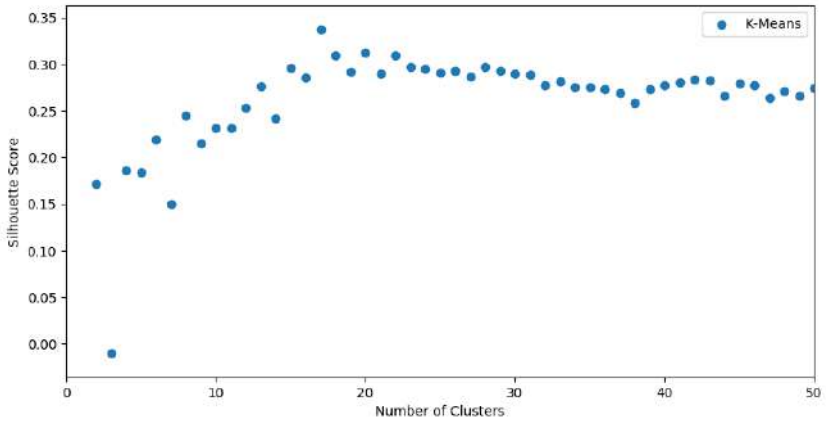


FIGURE 5.2: Dependence of silhouette score on the number of clusters (K-Means). Source: Author.

From Fig. 5.1 and Fig. 5.2 the exact number of clusters is not obvious. However, it is possible to assume that it is near 20. The silhouette score value is not high, meaning that the clusters are not separable entirely (the best value is 1, the worst is -1, values near 0 indicate overlapping clusters - *Sklearn silhouette score documentation*). I chose the most promising variants and build the distribution of customers within clusters (17, 20, 21 clusters) (Fig. 5.3).

K-Means performs extremely fast compared to all other algorithms used in the research. A customer distribution within the clusters is relatively good. These three numbers of clusters are added to further Markov chain modeling.

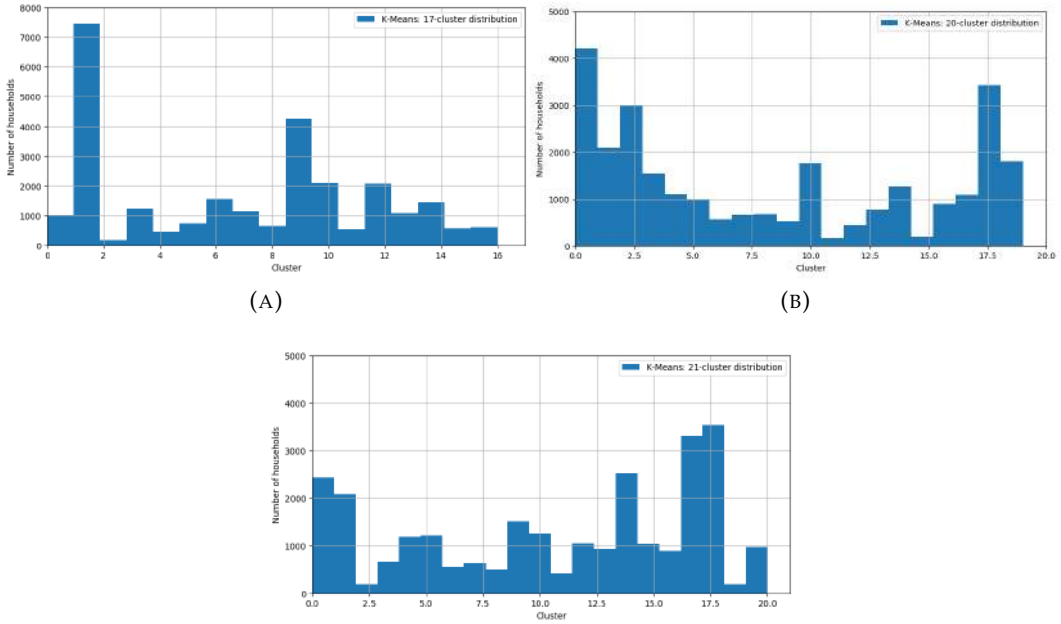


FIGURE 5.3: Histograms of customer distribution within 17(A), 20(B), and 21 (C) clusters (K-Means). Source: Author.

5.2.2 Gaussian Mixture Model

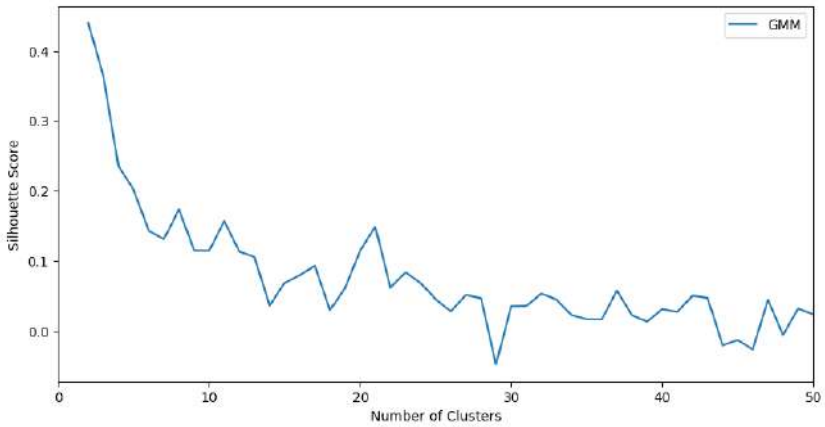


FIGURE 5.4: Dependence of silhouette score on the number of clusters (GMM). Source: Author.

According to the GMM silhouette score metric (Fig. 5.4), the most appropriate number of clusters are 8, 11 and 21. However, 17 and 20 also have a high value compared to their neighbors.

The chosen number of GMM clusters (Fig. 5.5) showed a less balanced customer distribution compared to K-Means results. For further investigation, the same number of clusters as for K-Means (17, 20 and 21) was selected.

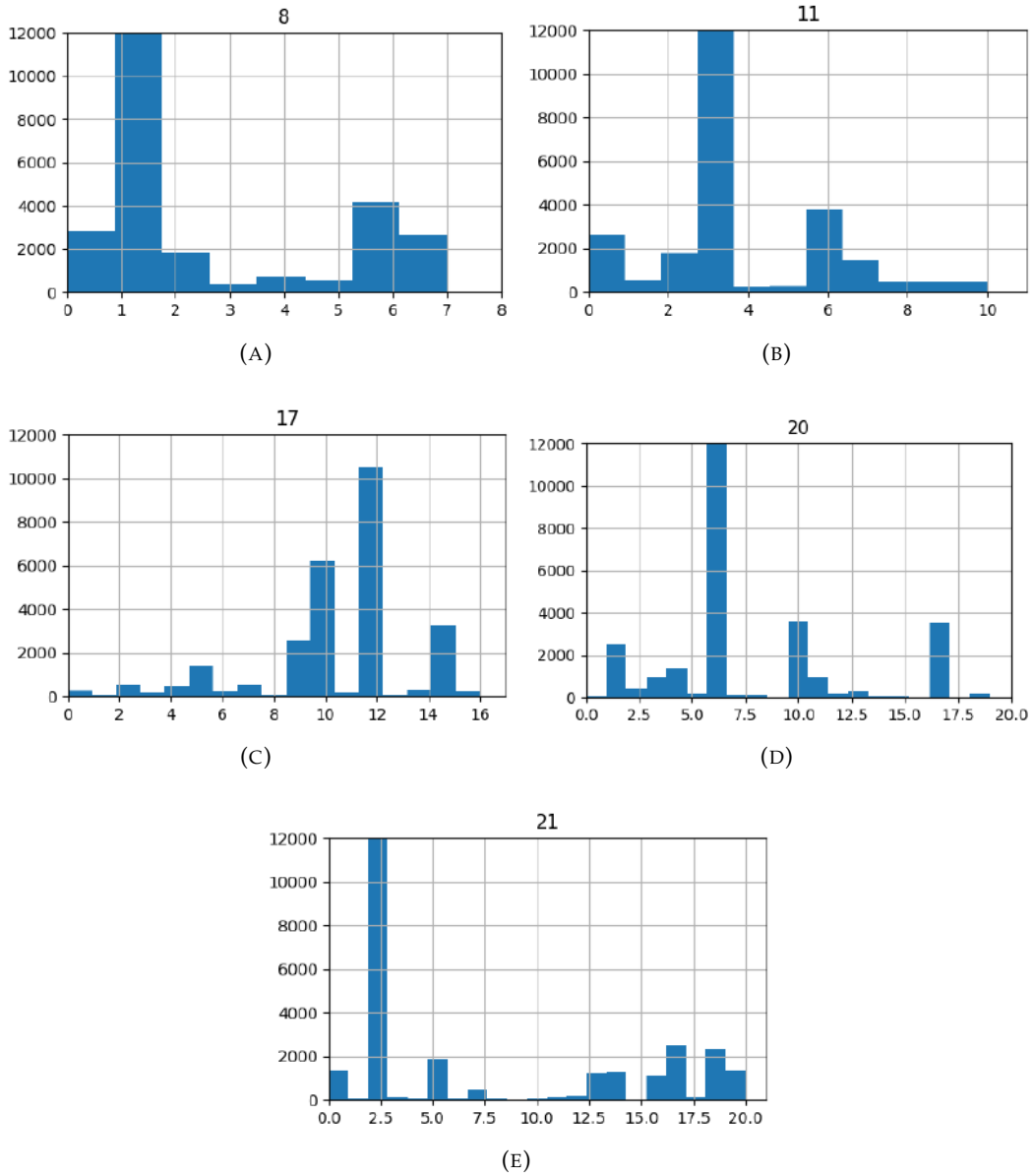


FIGURE 5.5: Histograms of customer distribution within 8 (A), 11 (B), 17 (C), 20 (D), 21 (E) clusters (GMM). Source: Author.

5.2.3 DBSCAN

The application of DBSCAN clustering was not so successful as K-Means or GMM. Although a number of clusters are not required, there are still two parameters to specify: a distance which indicates how close data points should be to each other to be considered as a part of a cluster (ε) and a minimum number of data points to be considered as a cluster ($minPts$).

According to Sander et al., 1998 a number of $minPts$ should be set up as large as a double number of dimensions (dim) in the dataset $minPts = 2\,dim$, and if the dimensionality is high, this number can be increased to get better results. The value of ε should be chosen as small as possible. One of the methods is to use a k-distance graph, plotting the distance to the $k = minPts - 1$ nearest neighbor ordered from the largest to the lowest value. Appropriate values of are considered where the plot

shows an "elbow". If ε is too low, a large part of the data will be out of clusters; whereas if ε is high, clusters will merge and the majority of objects will be in the same cluster.

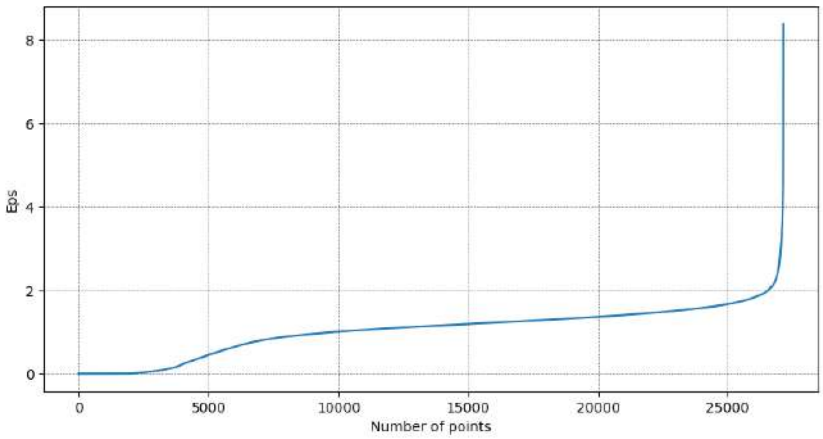


FIGURE 5.6: The Elbow method to find ε value for DBSCAN. Source: Author.

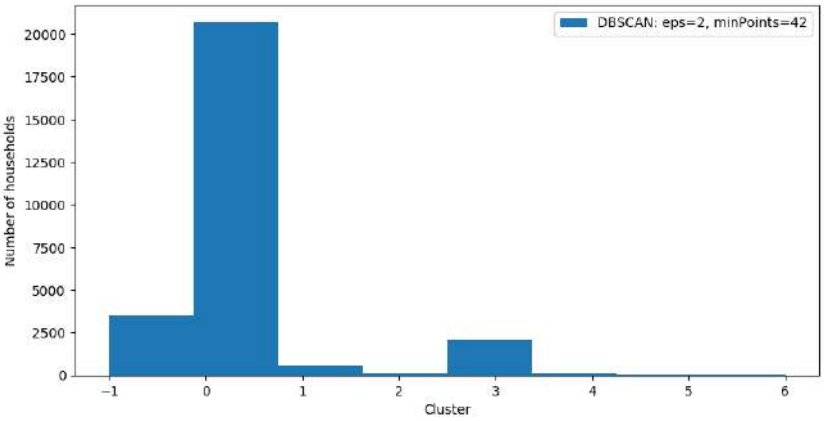


FIGURE 5.7: Histograms of customer distribution within 8 clusters (DBSCAN). Source: Author.

Since the dataset for clustering has 21 dimensions, *minPts* is considered to be equal to 42. From Fig. 5.6 it can be found that $\varepsilon = 2$. The results of applying DBSCAN algorithm are represented in Fig. 5.7. It detects 8 clusters (7 actual and 1 assigned for outliers) and shows that more than 20 thousand of households (about 75%) are assigned to one cluster and about 3.5 thousand households (about 13%) are considered as outliers.

Since the clustering with DBSCAN provides with small number of clusters and shows unbalanced household distribution among clusters (too many outliers and one strongly dominating cluster), it seems that this clustering technique is not appropriate for our dataset (DBSCAN tends to merge clusters with overlapping regions, and we know from GMM and K-Means that clusters are not entirely separated – have a low silhouette score) I have not included its results into further CLV modeling.

5.2.4 Summary

In this section, I ran multiple experiments with clustering algorithms (K-Means, GMM, and DBSCAN), figuring out the best one to fit the data. I refused to use DBSCAN results due to its poor performance on the data. I used a silhouette score for GMM and both Elbow Method (WCSS metric) and silhouette score for K-Means clustering to determine the optimal number of clusters. Moreover, I checked the distribution of customers in each clustering result. I decided to use the top three best results from K-Means and GMM clustering in further CLV modeling with Markov Chain.

5.3 Markov Chain

5.3.1 CLV modeling

At the beginning of this stage, I had the results from K-Means and GMM clustering (17, 20 and 21 clusters) – overall six different labels for each household in train and test datasets (5 thousand unique households are taken into account). Using a train dataset (first 20 4-week months) I calculated the transition matrix – a probability to move from one cluster to another (Fig. 5.8), average revenue and

$$r = (1 - churn\,rate) \tag{5.2}$$

for each cluster.

To model CLV for the selected period of the time (test number of months: eight 4-week months) I needed first to estimate revenue of each cluster (Fig. 5.9) using the following formula:

$$\hat{R}_{jn} = T^n < R_j > \tag{5.3}$$

where T – transition matrix, n – number of periods (months of interest), $< R_j >$– average revenue of a cluster j obtained for the past period (train dataset).

In the example with 21 clusters obtained by K-Means algorithm (Fig. 5.8-5.9) all customers who formed the 1st cluster are churners. Thus, the retention rate of the 1st cluster equals 0, and for all the rest clusters 1.

Then, CLV per cluster can be estimated as a sum of estimated revenues from now to a month of interest multiplied by the average retention rate value:

$$CLV_{jn} = \sum_{i=1}^{n} \hat{R}_{ji} < r_j > \tag{5.4}$$

where R_{ij} - estimated revenue for a cluster j in a month i, n – number of periods, $< r_j >$– average retention rate of a cluster j obtained for the past period (train dataset).

5.3.2 Model evaluation

To evaluate a model, a comparison with a ground truth should be performed. As ground truth, I used a test dataset where each household had six predicted labels (17, 20, 21 clusters' labels from K-Means and GMM) and calculated revenue per cluster for eight 4-weeks months (Fig. 5.10).

From Table 5.1 the best result (the lowest error) observed is for 21 clusters obtained by running K-Means algorithm. This result proves that the more balanced

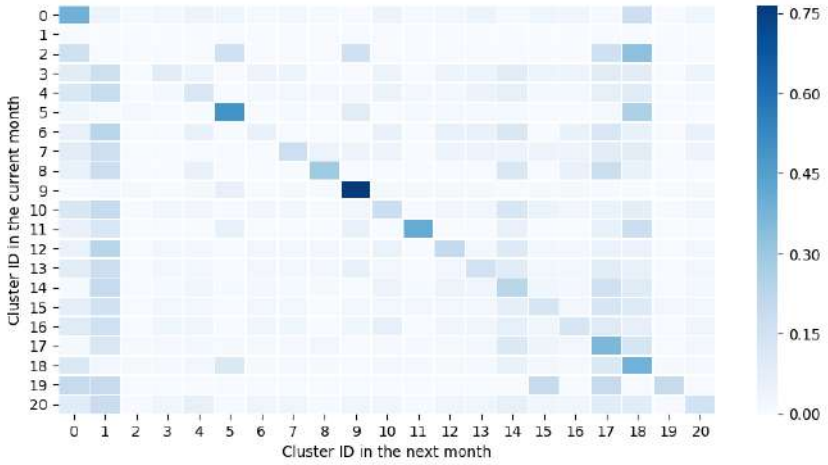


FIGURE 5.8: A sample of a Transition Matrix: probability to switch the cluster in the next month (21-clusters, K-Means). Source: Author.

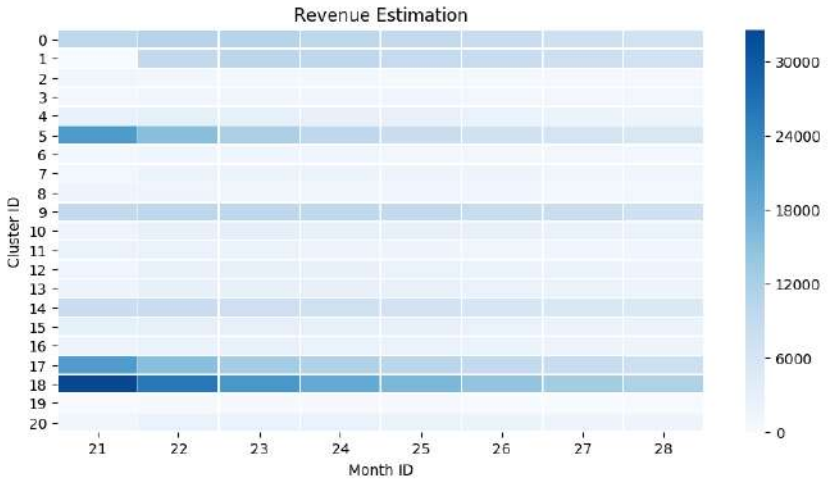


FIGURE 5.9: A sample of revenue estimation for the next 8 months (21-clusters, K-Means). Source: Author.

clusters the better result could be expected in further modeling. Thus, for survival analytics and time series forecasting, I used labels corresponding 21-cluster K-Means clustering result.

TABLE 5.1: Comparison of Markov Chain RMSE of CLV for different clustering techniques.

Method of clustering	**Number of clusters**	**RMSE**
K-Means	17	7140
	20	4987
	21	**4803**
GMM	17	7396
	20	6017
	21	6596

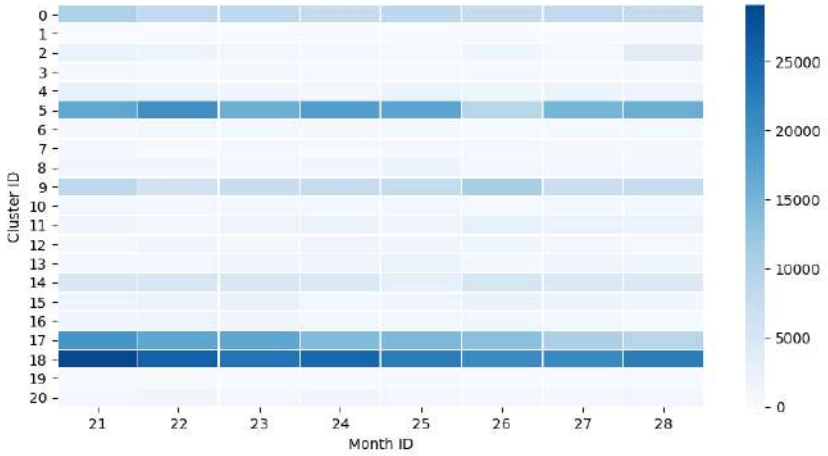


FIGURE 5.10: A sample of actual Revenue (ground truth for 21-clusters, K-Means). Source: Author.

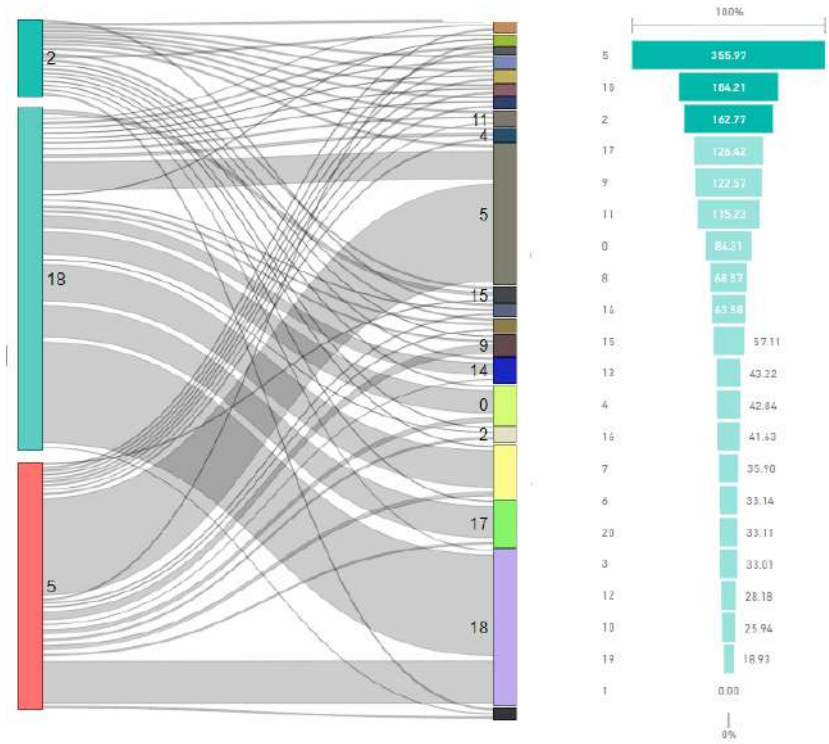


FIGURE 5.11: A sample of Sankey diagram of customer transitions between clusters. Source: Author.

5.3.3 Business Value vs CLV forecast

A stability of each cluster can be checked in Fig. 5.8. The diagonal elements of the matrix correspond to the probability to stay at the same cluster. Thus, clusters 0, 5, 9, 11, 17 and 18 show the highest percent of customers who perform the same purchase behavior from month to month. Moreover, it is easy to notice that the 1st cluster is formed by churners (customers who left the store this month). From the additional data exploration results, no other clusters contain churners. Thus, if a discount rate is neglected, the revenue of a cluster equals its CLV for a particular month.

Customer transitions between their states (belonging to a particular cluster in a

particular time period) can be represented as a Sankey diagram (Fig. 5.11). Such a representation is particularly valuable from the business perspective, since it displays the customer behavior in more understandable way compared to a transition matrix representation. Using this visualization, it is easy to distinguish the customer proportions between clusters and the way it was formed.

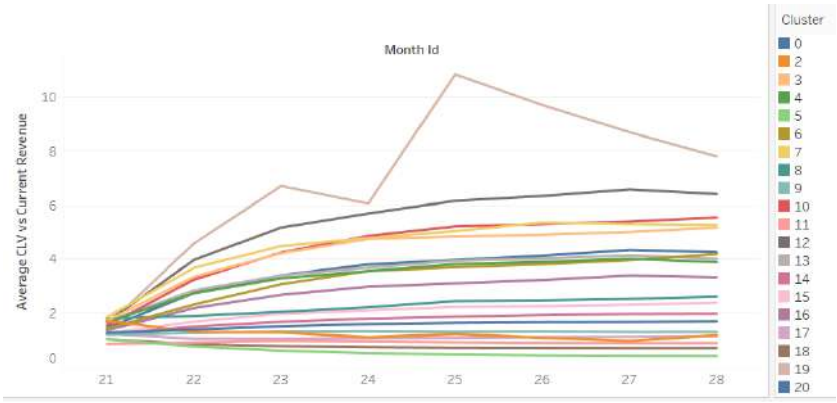


FIGURE 5.12: Average predicted CLV vs Current Revenue.
Source: Author.

Besides that, if to check predicted average CLV versus current revenue (Fig. 5.12) it is possible to see what actions towards customers could be performed. At the place where the considered ratio reaches maximum value the growth is expected, so some promotions for customers can be applied. At the place where the considered ratio decreases some recommendations towards a retention actions can be given. Thus, from Fig. 5.12 the most pessimistic scenarios are belonged to the 2nd, 5th and 18th clusters and the most optimistic ones to the 19th, 7th, 10th and 13th clusters.

5.3.4 Summary

In this section, the process of CLV estimation using Markov Chain model was considered. Six experiments were conducted for customer distributions between clusters obtained by K-Means and GMM clustering algorithms. The best RMSE was obtained for 21-cluster customer split by K-Means (the one which showed the most balanced customer distribution between clusters (Fig. 5.3)). Moreover, two visualizations of transitions between clusters and their practical application were considered too.

5.4 Survival Analytics

5.4.1 Cox's Proportional Hazard Model

To conduct survival analytics the train and test datasets were generated separately. The dataset included the first 20 4-week months of subsample data with five thousand unique households. The data was split with 80% for the train and 20% for the test. Cox's Proportional Hazard model summary is represented in Fig. 5.14. As we can notice (Fig. 5.13), there are those households who are ready to churn already and those who are going to stay with a store in a long relationship.

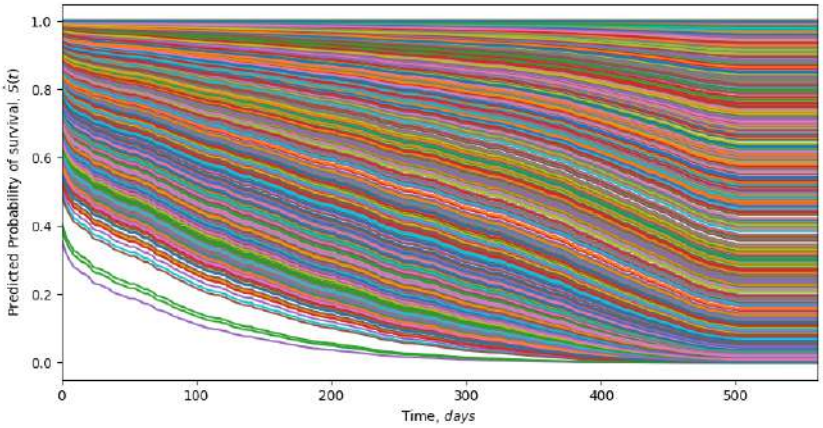


FIGURE 5.13: The time dependence of the probability to survive. Source: Author.

To obtain the lifespan which can be used together with time series revenue predictions, the CPH model (Fig. 5.14) was used to predict the behavior of the customers starting from the 21st 4-week month. The predicted lifespan was corresponded to the whole period of relationships without linkage to the calendar dates, to know how many days left for a particular household it is needed to subtract the number of days from the first purchase to the date the modeling was conducted. A threshold of survival probability was set as 0.5 (however, it can be set to a higher value to be more precise) to obtain the lifespan of the relationship between a store and a particular household.

5.4.2 Accuracy and calibration of CPH model

The concordance index (a fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered, **Raykar**) of the CPH model is 85%.

From the calibration curve above (Fig. 5.15) it can seen that the model has the best performance when predicting lifespan is located between 300 and 400 days. As examples, let's check the calibration plots for lifespan at 100 and 450 days (Fig. 5.16) – examining various probabilities against fractions that are present in the test dataset. The diagonal line represents perfect calibration. The model underpredicts the risk to churn for lifespans equals 100. For lifespan equals 450, the model slightly underpredicts risk to churn at the probability of less than 20% and overpredict risk to churn at the probability of more than 20%.

From Fig. 5.15 and Fig. 5.16 it can be deriven that the longer the customer stays with a company, the better churn prediction is. However, it is worth mentioning that the train dataset contained 80 weeks (560 days) date range only, so the maximum possible value of the lifespan was 560. If to check the predicted values of lifespan, it can be noticed that 65% of them were predicted as infinity (means that these customers are loyal and not going to churn from the store) and replaced with lifespan equals 600 for visualization (Fig. 5.17).

5.4.3 CPH CLV Benchmark model

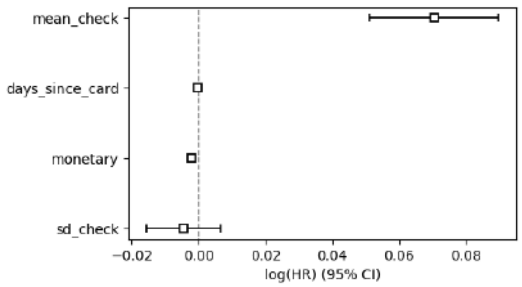
Preliminary I was going to use the result of CPH modeling together with Time Series predictions only. However, CLV can also be roughly estimated as a multiplication of

model	lifelines.CoxPHFitter
duration col	'duration'
event col	'churn'
number of observations	3315
number of events observed	2317
partial log-likelihood	-16186.96
time fit was run	2020-01-07 19:27:06 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
monetary	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-27.00	<0.005	531.08
mean_check	0.07	1.07	0.01	0.05	0.09	1.05	1.09	7.13	<0.005	39.83
sd_check	-0.00	1.00	0.01	-0.02	0.01	0.98	1.01	-0.79	0.43	1.22
days_since_card	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-17.99	<0.005	237.99

Concordance	0.85
Log-likelihood ratio test	2379.50 on 4 df, -log2(p)=inf

(A)



(B)

FIGURE 5.14: Cox's Proportional Hazard model summary. Source: Author.

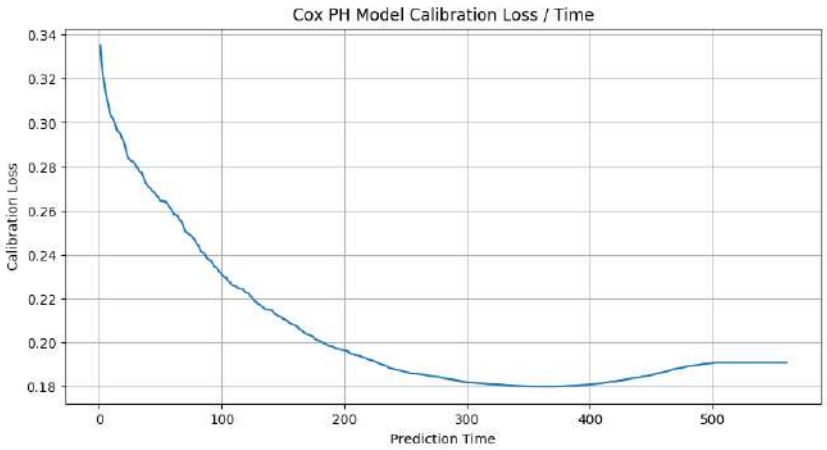
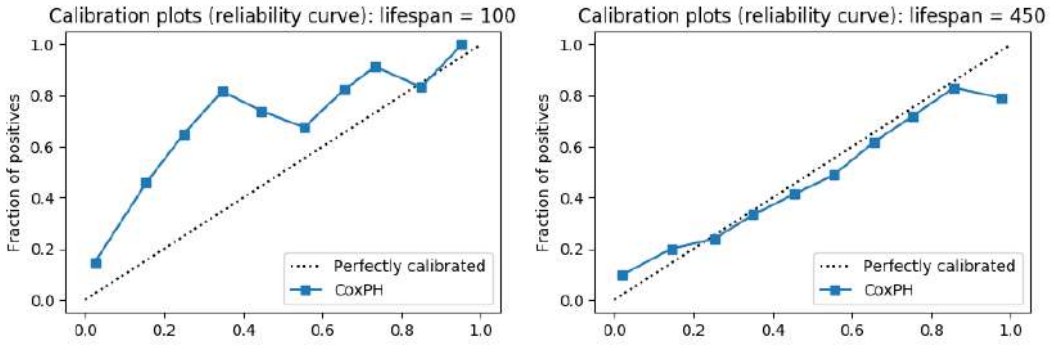


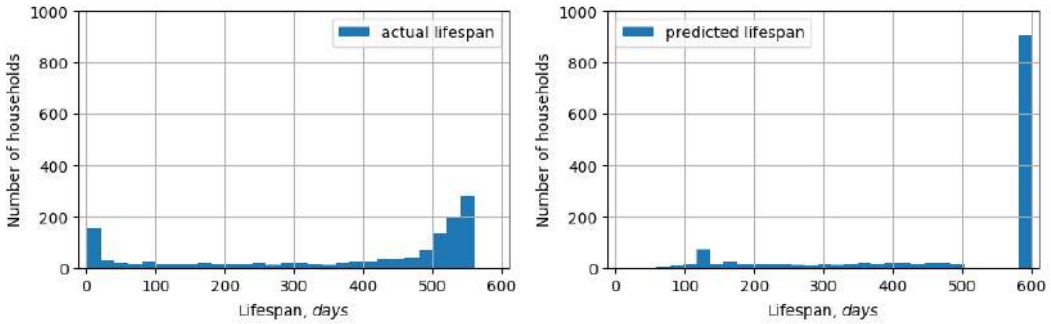
FIGURE 5.15: Cox's Proportional Hazard Model calibration loss over lifespan. Source: Author.

the average amount of money spent per a time unit (derived from the past transactions) and the number of time units (predicted by CPH model) and used as a benchmark model. To be consistent with Markov Chain model modeling prerequisites, the dataset used in the study contained the same subsample of unique households and the same date range. To validate the model, predicted values of revenue of each household for 21-28th 4-week months were compared with actual monthly revenue values. The error of the CLV model (RMSE) is 274. The drawback of this method is



(A) (B)

FIGURE 5.16: Example of Calibration plots for the lifespan of 100 (A) and 450 (B) days. Source: Author.



(A) (B)

FIGURE 5.17: Customer distribution by lifespan: actual versus predicted. Source: Author.

that future customers are eliminated from the CLV modeling.

5.4.4 Summary

In this section, Cox' Proportional Hazard Model was explored. By running the model, the lifespan for the customers was predicted. However, to be able to use it in further calculations, a threshold for the probability to survive was required to be set, and the number of days since customer's first transaction should be subtracted from the predicted values of lifespans. Then the model performance was checked by exploring calibration curve. In addition, Cox' Proportional Hazard CLV Benchmark model was built.

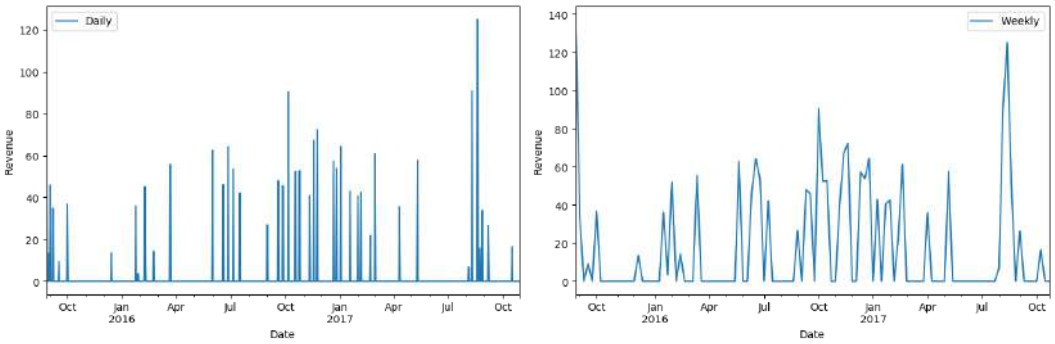
5.5 Time Series Forecasting

For the time series forecasting, I used data with labels for 21 clusters obtained by the K-Means algorithm. There are multiple approaches on how to aggregate the data to proceed with the revenue forecast: aggregation on household and cluster level on a daily, weekly and monthly (4-week month) basis.

A monthly aggregation approach for both individual and cluster level was excluded from the study at the very beginning due to too narrow date range in the

available dataset (assuming that 20 points are not enough to build a good time series prediction model according to Groebner et al., 2010 a sufficiently large sample is equal at least 30).

After the visualization of timeseries for a randomly chosen household on a daily and weekly basis (Fig. 5.18), these two approaches were excluded from the research too and left for later exploration with the methods applicable for the intermittent data forecasting (Croston's and Bootstrapping methods, Teunter and Duncan, 2009).



(A) (B)

FIGURE 5.18: Revenue from a randomly chosen household with daily (A) and weekly (B) aggregations. Source: Author.

5.5.1 Weekly Time Series on Cluster Level

ARIMA, SARIMA

Let us consider as an example on of the promising clusters from the Markov Chain modeling (Fig. 5.9-5.10), the 9th cluster. Its weekly revenue over time is represented as a time series dependence on Fig. 5.19. To be consistent with Markov Chain's assumptions the first twenty 4-week months of data (80 weeks) was used for the time series modeling and the rest eight 4-week months data (32 weeks) was kept for validation of the forecast with overall randomly selected 5 thousand of unique households.

A Dicky-Fuller test showed that the first order difference was enough for data preprocessing to obtain stationary data (ADF-statistics: -4.38, p-value: 0.0003). The

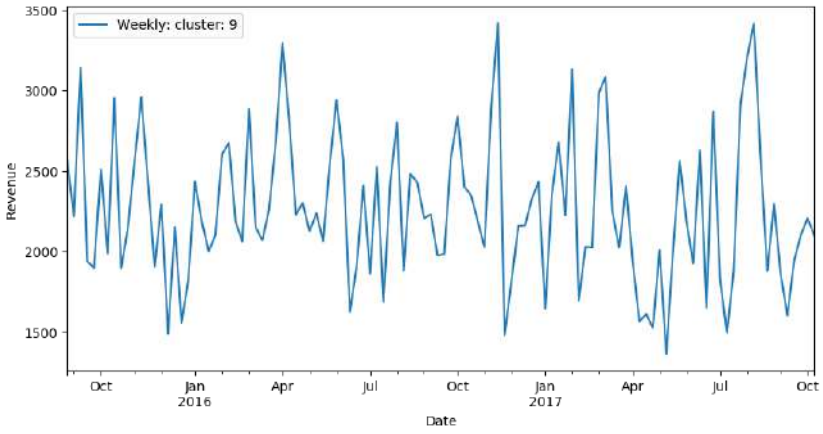
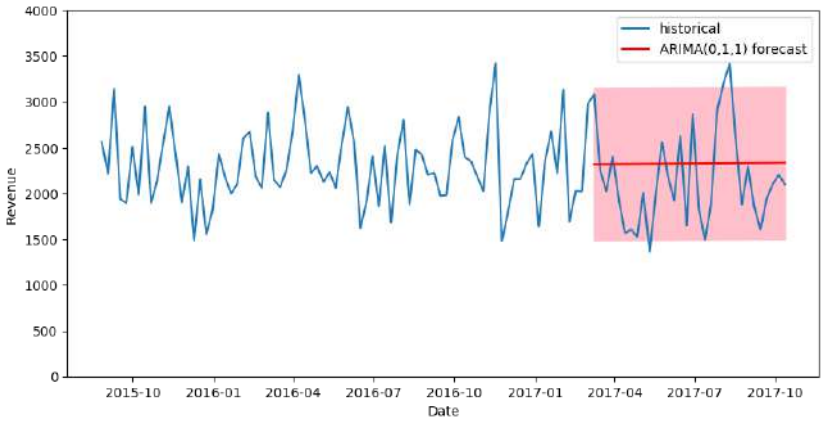
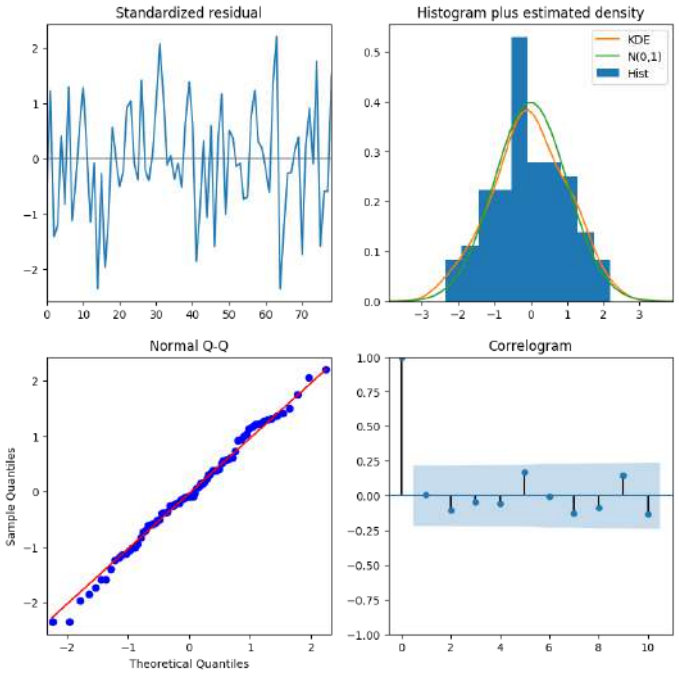


FIGURE 5.19: Revenue Time Series of the 9th cluster. Source: Author.

evidence of the seasonality was not found by inspecting ACF and PACF plots. Different ARIMA models were compared by AIC metric. The forecast of the model ARIMA(0,1,1) with the minimum AIC=1189.5 and its residuals analysis is represented in Fig. 5.20. According to the forecast (Fig. 5.20), there is an expected slight increase of revenue for about 20 dollars per this cluster during the next 32 weeks ($<$ 1% of total revenue), which is not far from the average revenue per cluster used in Markov Chain modeling assumptions and CPH benchmark model. The error of the Time Series model for the 9th cluster (RMSE) is 518.



(A)



(B)

FIGURE 5.20: Actual vs Forecasted Revenue of the 9th cluster and model error analysis. Source: Author.

5.5.2 Business Value vs CLV forecast

Following this approach of data analysis, there is also a possibility to have other useful for business insights about customers' behavior. If to represent the revenue

as an average per cluster with its forecast (Fig. 5.21) and plot CPH model curves for each household who form the cluster (Fig. 5.22), the average expected revenue from a single household and its stability over time in terms of lifespan can be checked. As for the 9th cluster, the average value of revenue per household fluctuates around 30 dollars and customers of the cluster are expected to stay with a company with a probability of 50% for at least 250 days.

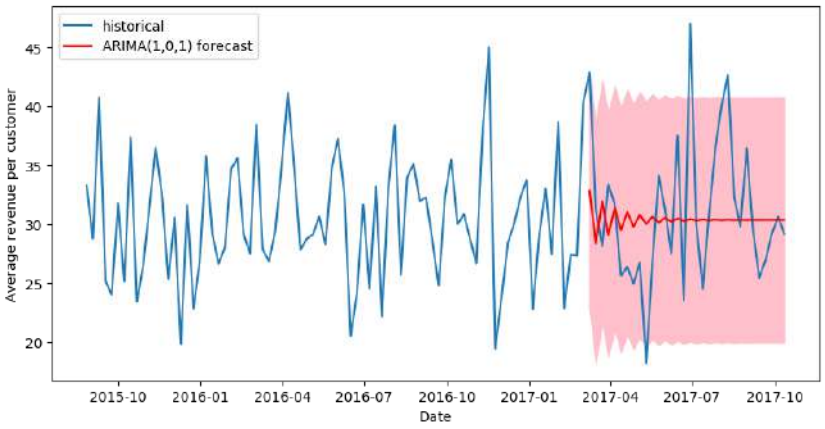


FIGURE 5.21: Average revenue per customer of the 9th cluster. Source: Author.

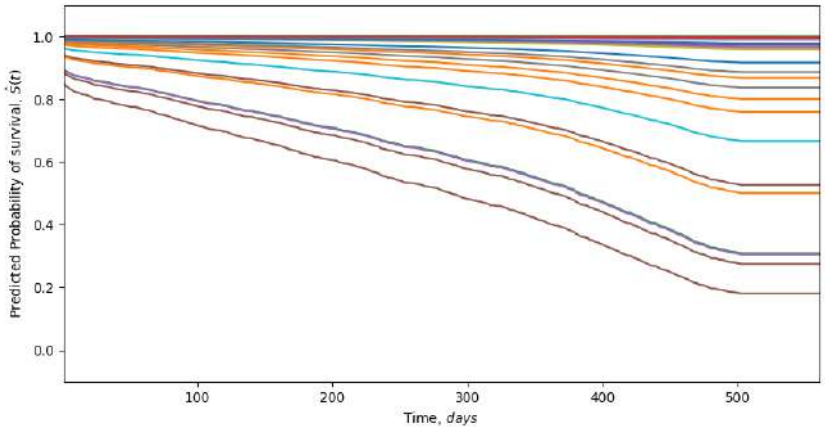


FIGURE 5.22: Probability to survive for the households of the 9th cluster: general and at the point of lifespan prediction. Source: Author.

A similar conclusion can be derived from the Markov Chain probability matrix (Fig. 5.8). The 9th cluster performs relatively stable towards churners of the store as well as switchers to the other clusters.

5.5.3 Summary

In this section, a time series approach for revenue estimates was checked on individual and cluster level. As an example of the model results the 9th cluster. The output of the model together with the results of Cox' Proportional Hazard model (lifespan) was used in CLV calculations. Moreover, it was shown how to use both methods to extract other valuable information for the business. To be able to compare the

performance of the model with Markov Chain model output, the estimation of CLV was done on weekly aggregated data and then mapped to months.

5.6 Summary

In this research, three different clusterization techniques (K-Means, GMM, DBSCAN) were examined towards a performance on the transactional and loyalty card data with a chosen feature set. K-Means has shown the best result in terms of speed and the quality of customer distribution in clusters. The output of the clustering approaches was later used in Markov Chain, Survival Analytics, and Time Series CLV modeling.

The fundamental part of the Markov Chain model used for CLV modeling was a state definition. The states represented by customer clusters had three main components: a retention rate, a revenue and a behavioral pattern (probability to move between clusters). Markov Chain CLV estimates were modeled based on transition probabilities (derived from the historical transactional data), average cluster monthly revenues realized within the first 20 4-week months and an average retention rate per cluster

On the other hand, econometric Cox' Proportional Hazard model estimated a customer probability to churn in a particular period taken into account the factors which could cause such an event. Time Series Forecasting techniques estimated the value of revenue for future periods which was approximately equal to an average revenue per cluster obtained in the past. The results of both models were combined in CLV estimates.

Finally, when a particular cluster was considered within results of both models, the general picture obtained by a probabilistic approach on a cluster level was proven by econometric approach on an individual one.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

CLV is an important metric for retail since it indicates the success of the marketing campaigns and channels, deals with retention and churn, and helps with decisions towards resource allocation. However, the modeling of CLV for retail is a challenging task due to the lack of access to the historical data of purchases and difficulty of customer identification.

In this research, an analytical framework for offline and semi-offline businesses was prototyped (Kasprova, 2020: *CLV for Retail*) to estimate CLV based on transactional and loyalty card data . This framework transfers the raw data obtained from the grocery store operational databases to business insights based on CLV estimates and their visualizations. It can be used for channels and campaign evaluations.

According to the data pipeline of the framework, the transactional data of the store linked to the loyalty card data provides historical references of purchases to corresponding customers. Then customers are segmented based on the similarity of their purchase behavior performed by clusterization algorithms. Markov Chain probabilistic model applied to segments of customers shows a general picture of the ongoing processes which are easy to interpret exploring the Sankey diagram. To have more details on an individual level to transfer the previously received information into data-driven decisions or possible actions towards CRM and resources allocation Time Series revenue forecasts, together with Cox's Proportional Hazard model, lifetime estimates are used.

There are a few limitations of the research that should be highlighted. The proposed approaches of CLV estimations were developed for existing customers only. The framework was developed using 1.5-year data and tested only on 0.5-year data. It can be significantly improved using a dataset with the broader data range, at least to be able to make monthly based aggregations for the Time Series predictions. The aim of the research and a framework development had a more conceptual character rather than accuracy-oriented. Thus, there is a variety of methods, which can be explored for a more efficient revenue and churn prediction in the future.

6.2 Future Work

1. To check other time series analysis approaches: classical - Facebook Prophet, VARMAX, Holz-Winter, and RNN – LSTM, seq2seq, ES-RNN, DeepAR etc.

2. To explore ARIMAX/SARIMAX models to be able to check the performance of multivariate time series and compare the obtained results with my current state of art model, including public holidays, school vacation, store location, and festivals nearby, shop holidays, fasting and abstinence as exogenous variables.

3. To explore time series approaches to predict revenue on customer level when the data is intermittent: Croston's and Bootstrapping methods (Teunter and Duncan, 2009).

4. To predict churn using machine learning approaches such as logistic regression, binary classification with XGBoost, SVM, Random Forest, etc.

5. To segment customers based on their revenue and then include their taste features.

6. To extend modeling to other stores.

7. To develop interactive visualization of results in Dash (a visualization framework with a Python backend).

Bibliography

Batislam E.and Denizel, M. and A Filiztekin (2007). "Empirical Validation and Comparison of Models for Customer Base Analysis". In: *International Journal of Research in Marketing* 24.3, pp. 201–209.

Bishop, Ch. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Brownlee, J. (2018). *Introduction to Time Series Forecasting with Python*. v.1.4, e-book.

Cox, D. (1972). "Regression Models and Life Table". In: *Journal of the Royal Statistical Society, Series B* 34.2, pp. 187–220.

Damodaran, A. (2020). *Cost of Capital by Sector (US)*. URL: http://people.stern.nyu.edu/adamodar/New_Home_Page/datafile/wacc.html. Accessed: 2020-01-08.

Ester, M. et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland*, pp. 226–231.

Estrella-Ramón, A. et al. (2013). "A marketing view of the customer value: Customer lifetime value and customer equity". In: *South African Journal of Business Management* 44.4, 47–64.

Forecasting using ARIMA models in Python. URL: https://www.datacamp.com/courses/forecasting-using-arima-models-in-python. Accessed: 2020-01-08.

Fox, E. and C. Guestrin (2016). *Gaussian Mixture Model for Clustering*. https://www.youtube.com/watch?v=DODphRRL79c. Accessed: 2020-01-08.

Groebner, D. et al. (2010). *Business Statistic*. 8th edn, Pearson.

Gupta, S. et al. (2006). "Modeling Customer Lifetime Value". In: *Journal of Service Research* 9, pp. 139–155.

Haenlein, M., A. Kaplan, and A. Beeser (2007). "A Model to Determine Customer Lifetime Value in a Retail Banking Context". In: *European Management Journal* 25.3, 221–234.

Hughes, A. (2011). *Strategic Database Marketing*. 4th edn.,McGraw-Hill, New York.

Jasek, P. et al. (2018). "Modeling and Application of Customer Lifetime Value in Online Retail". In: *Informatics*.

Jasek, P. et al. (2019). "Predictive performance of customer lifetime value models in e-commerce and the use of non-financial data". In: *Prague Economic Papers*.

K-Means and X-Means Clustering. https://www.brandidea.com/kmeans.html. Accessed: 2020-01-08.

Kaplan, L. and P. Meier (1958). "Nonparametric Estimation form Incomplete Observations". In: *Journal of American Statistical Association* 53.6, 457–481.

Kasprova, A. *CLV for Retail*. https://github.com/kasprova/advanced_customer_analytics. Accessed: 2020-01-08.

Kotler, P. (1974). "Marketing During Periods of Shortage". In: *Journal of Marketing* 38.3, 20–29.

Lloyd, S. (1982). "Least square quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137.

Machine Learning for Marketing Analytics in R. https://campus.datacamp.com/courses/marketing-analytics-in-r-statistical-modeling. Accessed: 2020-01-08.

McKinsey. *Customer Lifecycle Management*. https://www.mckinsey.com/business-functions/marketing-and-sales/how-we-help-clients/customer-lifecycle-management. Accessed: 2020-01-08.

Miller, R. (2011). *Survival analysis*. 2nd edn., Wiley.

Mittal, A. (2019). *Understanding RNN and LSTM*. URL: https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e. Accessed: 2020-01-08.

Neslin, S. et al. (2006). "Defection detection: Measuring and understanding the predictive accuracy of customer churn models". In: *Journal of Marketing Research* 46.5, pp. 204–211.

Nikkhahan, B., Badrabadi A. Habibi, and M Tarokh (2011). "Customer Lifetime Value model in an online toy store". In: *Journal of Industrial Engineering International* 7.12, pp. 19–31.

Overview of clustering methods. https://scikit-learn.org/stable/modules/clustering.html. Accessed: 2020-01-08.

Pfeifer, P. and R. Carraway (2000). "Modelling customer relationships as Markov chains". In: *Journal of Interactive Marketing* 14.2, pp. 43–55.

Sander, J. et al. (1998). "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications". In: *Data Mining and Knowledge Discovery 2* 2, 169–194.

Schubert, E. et al. (2017). "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN". In: *ACM Transactions on Database Systems (TODS)* 42.3, p. 19.

Seabold, Skipper, and Josef. Perktold (2010). "Statsmodels: Econometric and statistical modeling with python". In: *Proceedings of the 9th Python in Science Conference*.

Sklearn silhouette score documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. Accessed: 2020-01-08.

Teunter, R. and L. Duncan (2009). "Forecasting intermittent demand: a comparative study". In: *Journal of the Operational Research Society* 60.3, pp. 321–329.

Villanueva, J. and D. Hanssens (2007). "Customer equity: Measurement, management and research opportunities". In: *Foundations and Trends in Marketing* 1.1, pp. 1–95.

Wedel, M. and W. Kamakura (1999). *Market Segmentation: Conceptual and Methodological Foundations*. 2nd edn. Kluwer Academic Publishers, Boston.