KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Computer Science and Engineering, specializing in Machine Learning

Second cycle, 30 credits

Shoppin' in the Rain

An Evaluation of the Usefulness of Weather-Based Features for an ML Ranking Model in the Setting of Children's Clothing Online Retailing

ISAC LORENTZ

Stockholm, Sweden, 2023

Shoppin’ in the Rain

An Evaluation of the Usefulness of Weather-Based Features for an ML Ranking Model in the Setting of Children’s Clothing Online Retailing

ISAC LORENTZ

Date: September 17, 2023

Supervisors: Ahmad Al-Shishtawy, Marcus Svensson
Examiner: Danica Kragic Jensfelt
School of Electrical Engineering and Computer Science
Host company: Babyshop Sthlm AB
Swedish title: Handla i regnet
Swedish subtitle: En utvärdering av användbarheten av väderbaserade variabler för en ML-rankningsmodell inom onlineförsäljning av barnkläder

Abstract

Online shopping offers numerous benefits, but large product catalogs make it difficult for shoppers to understand the existence and characteristics of every item for sale. To simplify the decision-making process, online retailers use ranking models to recommend products relevant to each individual user. Contextual user data, such as location, time, or local weather conditions, can serve as valuable features for ranking models, enabling personalized real-time recommendations. Little research has been published on the usefulness of weather-based features for ranking models in online clothing retailing, which makes additional research into this topic worthwhile. Using Swedish sales and customer data from Babyshop, an online retailer of children's fashion, this study examined possible correlations between local weather data and sales. This was done by comparing differences in daily weather and differences in daily shares of sold items per clothing category for two cities: Stockholm and Göteborg. With Malmö as an additional city, historical observational weather data from one location each in the three cities Stockholm, Göteborg, and Malmö was then featurized and used along with the customers' postal towns, sales features, and sales trend features to train and evaluate the ranking relevancy of a gradient boosted decision trees learning to rank LightGBM ranking model with weather features. The ranking relevancy was compared against a LightGBM baseline that omitted the weather features and a naive baseline: a popularity-based ranker. Several possible correlations between a clothing category such as shorts, rainwear, shell jackets, winter wear, and a weather variable such as feels-like temperature, solar energy, wind speed, precipitation, snow, and snow depth were found. Evaluation of the ranking relevancy was done using the mean reciprocal rank and the mean average precision @ 10 on a small dataset consisting only of customer data from the postal towns Stockholm, Göteborg, and Malmö and also on a larger dataset where customers in postal towns from larger geographical areas had their home locations approximated as Stockholm, Göteborg or Malmö. The LightGBM rankers beat the naive baseline in three out of four configurations, and the ranker with weather features outperformed the LightGBM baseline by 1.1 to 2.2 percent across all configurations. The findings can potentially help online clothing retailers create more relevant product recommendations.

Keywords

Statistical analysis, regression analysis, recommender systems, ensemble learning, electronic commerce, LightGBM, learning to rank, feature selection, weather-based features, fashion

Sammanfattning

Internethandel erbjuder flera fördelar, men stora produktsortiment gör det svårt för konsumenter att känna till existensen av och egenskaperna hos alla produkter som saluförs. För att förenkla beslutsprocessen så använder internethandlare rankningsmodeller för att rekommendera relevanta produkter till varje enskild användare. Kontextuell användardata såsom tid på dygnet, användarens plats eller lokalt väder kan vara värdefulla variabler för rankningsmodeller då det möjliggör personaliserade realtidsrekommendationer. Det finns inte mycket publicerad forskning inom nyttan av väderbaserade variabler för produktrekommendationssystem inom internethandel av kläder, vilket gör ytterligare studier inom detta område intressant. Med hjälp av svensk försäljnings- och kunddata från Babyshop, en internethandel för barnkläder så undersökte denna studie möjliga korrelationer mellan lokal väderdata och försäljning. Detta gjordes genom att jämföra skillnaderna i dagligt väder och skillnaderna i dagliga andelar av sålda artiklar per klädeskategori för två städer: Stockholm och Göteborg. Med Malmö som ytterligare en stad så gjordes historiska metereologiska observationer från en plats var i Stockholm, Göteborg och Malmö till variabler och användes tillsammans med kundernas postorter, försäljningsvariabler och variabler för försäljningstrender för att träna och utvärdera rankningsrelevansen hos en gradient-boosted decision trees learning to rank LightGBM rankningsmodell med vädervariabler. Rankningsrelevansen jämfördes mot en LightGBM baslinjesmodel som saknade vädervariabler samt en naiv baslinje: en popularitetsbaserad rankningsmodell. Flera möjliga korrelationer mellan en klädeskategori som shorts, regnkläder, skaljackor, vinterkläder och och en daglig vädervariabel som känns-som-temperatur, solenergi, vindhastighet, nederbörd, snö och snödjup upptäcktes. Utvärderingen av rankingsrelevansen utfördes med mean reciprocal rank och mean average precision @ 10 på ett mindre dataset som bestod endast av kunddata från postorterna Stockholm, Göteborg och Malmö och även på ett större dataset där kunder med postorter från större geografiska områden fick sina hemorter approximerade som Stockholm, Göteborg eller Malmö. LigthGBM-rankningsmodellerna slog den naiva baslinjen i tre av fyra konfigurationer och rankningsmodellen med vädervariabler slog LightGBM baslinjen med 1.1 till 2.2 procent i alla konfigurationer. Resultaten kan potentiellt hjälpa internethandlare inom mode att skapa bättre produktrekommendationssystem.

Nyckelord

Statistisk analys, regressionsanalys, rekommendationssystem, ensemble-inlärning, näthandel, LightGBM, learning to rank, variabelselektion, väderbaserade variabler, mode

Acknowledgments

First and foremost, I would like to express my sincere gratitude to Babyshop for providing the data and computing resources for this thesis. I would also like to extend my thanks to my company supervisor Marcus Svensson, and my assistant company supervisor Adrián Campoy Rodríguez for all the help and guidance with the thesis and for everything that I learned along the way. Marcus calculated the sales trend features for the Babyshop data. I would also like to thank Einar Thor Gunnlaugsson, Hongyu He, and Tatyana Boronenkova at Babyshop for providing useful input to the thesis, helping me with SQL and the Babyshop Data Warehouse, and cheering on my progress. I would like to extend my thanks to my academic supervisor Ahmad Al-Shishtawy for his support and guidance throughout this degree project. I would also like to thank my examiner Danica Kragic Jensfelt. I would like to thank my classmates Wai-Hong Anton Fu, George Malki, Samuel Skoog, and Kunal Bhatanagar for knowledge sharing and help with proofreading and the general thesis process. Lastly, I would also like to thank Charlotta Lorentz for her help with proofreading.

Stockholm, September 2023
Isac Lorentz

Contents

List of Figures

List of Tables

Listings

List of Acronyms and Abbreviations

GBDT	Gradient Boosting Decision Trees
LETOR	Learning to Rank
MAP	Mean Average Precision
MAP@K	Mean Average Precision at K
ML	Machine Learning
MRR	Mean Reciprocal Rank
SMA	Simple Moving Average

Chapter 1

Introduction

1.1 Background

Online shopping grows more common every year, with 2 billion individuals having bought goods or services online in the year 2021 [1] and is for the year 2023 estimated to make up over 20 percent of global retail sales [2]. Benefits of shopping online are numerous, including flexibility, better prices, no need to physically go to the store, large product assortments, no crowds, and the ability to shop outside one's local geographical area [3]. The low marginal cost for online retailers for warehousing additional products enables online retailers to have large product assortments, including niche products that sell significantly less than the popular products [4, p.273]. Large product assortments allow consumers to find products that better suit their specific preferences and needs, but the large size of the typical online product catalog makes it difficult to get an overview of the full catalog and to know the existence and characteristics of each individual item. To aid online shoppers in combating this problem, e-tailers can use product ranking models and product recommender systems to help online customers find the products that they need and to recommend the products that are most relevant to the consumers. Such systems can include information about the user, the products, and the context of the user. Contextual information such as location, user intent, time, and sales channel can provide helpful information for recommender systems.

Millions of people globally can afford costly consumption of clothing and footwear for reasons such as expressing their personal style, belonging to certain groups, or showing their status. This means that clothing many times is not always bought strictly out of necessity. At the same time, people need different types of clothing to provide warmth and protection from the elements.

Seasonal variability of the weather causes individuals living in most parts of the world to use different types of clothing throughout the year to adapt to different weather conditions. This requires planning and shopping in order to own clothing that meets the varying demands of the different seasons.

As concluded by the World Meteorological Organization, climate change is currently aggravating local climate variability, causing larger variations in local weather [5, p.3]. This forces some businesses to adapt in order to continue to meet their customers' needs in the best way. For fashion retailing, this means that stores can potentially expect demand for unseasonal clothing. This is important for purchasing and supply chain management, but also for product recommendations for online stores.

The correlation between weather and purchasing patterns for adult clothing has been studied in several papers such as [6], [7], but little research has been done specifically on the correlation between weather and purchasing patterns for children's clothing. Shopping for adult's clothing and children's clothing differs in several ways. Adults typically buy most adult clothing for themselves, while children's clothing is typically bought for the child by her parent(s) or guardian(s). As children grow, they routinely need larger clothes. For these reasons, it could be reasonable to believe that correlations between weather and purchasing patterns for adult clothing cannot be fully extrapolated to the children's domain. Shopping for children's clothing could be associated with more planning or less spontaneity than shopping for adult clothing.

This thesis focused on the usefulness of using weather data as a type of contextual information for creating better product recommender systems and was conducted together with Babyshop, an e-tailer of children's clothing (among other item categories). Data professionals at Babyshop hold the hypothesis that local weather affects customer behavior when shopping for children's clothes and that this information can be used to provide better product recommendations.

1.2 Problem

1.2.1 Original problem and Definition

Online shoppers rely on search functions or recommender systems to browse large product catalogs when shopping for clothes online and to find the most relevant items that fit their shopping needs. Online retailers wish to serve their customers with the most relevant and personalized recommendations even in the case when limited user data is available. Weather information is part of a user's contextual information and can be used as a signal to model an online customer's purchasing intent. Local weather information has been known to correlate with retail sales of clothes (see section 2.4.1) and could possibly serve as a useful feature set in providing personalized shopping recommendations using an item ranking Machine Learning (ML)-model.

1.2.2 Research Question

How does the performance of an ML-based recommender system differ with or without weather features in the domain of ranking children's clothing in e-commerce?

1.3 Purpose

One purpose of this thesis is to gain insights into the usefulness of weather data as features for ML-based recommender systems in the task of recommending children's clothes. Perhaps, these findings could possibly be extrapolated to the domain of adult clothing. Useful findings could be of interest to the recommender systems community and in particular those interested in recommender systems in fashion, such as the FashionxRecsys community [8]. The Association for Computing Machinery holds recurrent conferences on Recommender Systems in Fashion & Retail with company presence from Zalando, ASOS, and Farfetch, among others [8].

If results indicate that weather data can be used to create better product recommendations, this could be used by Babyshop to potentially sell more clothes and create a better shopping experience on their shopping platforms. Better recommender systems could also help their customers to get their shopping needs better met.

The use of weather data as features for product recommendations for purchasing clothes is likely not new, but during the literature review, no published research on this exact topic was found. This makes additional research on this topic worthwhile. Several of the papers we found about weather-based clothing recommendations are in the domain of recommending clothes from a person's wardrobe, not from an online store, as in the case of this thesis.

1.4 Goals

There are two overarching goals of this project. The first goal is to explore possible correlations between local weather data and sales data on the Babyshop e-tailing platforms. The second is to train and evaluate a Gradient Boosting Decision Trees (GBDT) Learning to Rank (LETOR) (the application of machine learning for creating ranking models) recommender system that utilizes weather data as features and to compare it against a baseline that does not have weather features. These goals can be broken down into the following sub-goals:

Statistical Analysis

1. Collect weather data and Babyshop sales data
2. Develop a program for statistical analysis
3. Evaluate the statistical analysis and decide which weather features to use for the recommender system

Recommender systems

1. Collect Babyshop sales data for recommender systems
2. Develop a program for data processing, training and evaluation of recommender systems
3. Evaluate the recommender systems on the small dataset
4. Evaluate the recommender systems on the large dataset

1.5 Research Methodology

The first part of the thesis was the literature review. From studying previous work in clothing sales and weather correlations and recommender systems, useful insights regarding what methodology to use for this thesis could be obtained. Particularly by studying clothing sales and weather data correlations, information about which features are useful could be learned.

Studying the correlations between sales on the Babyshop platforms and local weather data using statistical analysis was partly done due to it being a request from Babyshop, but it also helps show what might be relevant weather features for the ML model. This way, the training time could be reduced since we did not introduce irrelevant features. The correlations tested for and the type of statistical analysis that was conducted were partly based on the findings in the related works.

The specific model used for the recommender system was LGBMRanker from the LightGBM Python library, but other potential choices for the ranking model include other GBDT libraries that also support LETOR such as CatBoost and XGBoost. LightGBM was selected based on several factors. The main reason for choosing LightGBM as the ranker model was that there are several high-quality Kaggle notebooks detailing solutions to a similar ranking problem, also in the fashion domain [9]. These notebooks were part of the H&M Personalized Fashion Recommendations competition [10] and served as tutorials in this project on how to use the LightGBMRanker model. This gave us confidence that the model could be implemented correctly in this project. The library has been used in winning solutions for several competitions for recommender systems [11]. The library handles categorical variables using a solution based on Fisher's approach for grouping based on minimized in-group-variance [12], which is a more optimal approach than one-hot-encoding for tree-based learners [13], [14]. The library is also popular with 15,000+ Github stars [15] and 8000+ citations on Google Scholar [16].

A constraint was that the recommender systems investigated in the thesis had to be able to be evaluated using past sales data from the Babyshop shopping platforms. This also restricted somewhat which data points in the dataset could be used since it required that the items in the evaluation data were also present in the training data. The data only has information about which products were bought and which products were not, so the data does not contain graded information about how relevant the products were to the users. This also means that evaluation metrics for recommender systems that consider the graded relevance of the products (such as normalized discounted cumulative gain) are not optimal for evaluation in this project. The final evaluation metrics were chosen since they are popular for the evaluation of recommender systems and they provide different information.

1.6 Delimitations

Sales data was restricted to children's clothing and shoes sold on the online storefronts operated by Babyshop. Weather data will only be taken from three cities in Sweden: Stockholm, Göteborg, and Malmö (from one location in each city).

Several delimitations were made in regard to the model architecture selection and the model training due to time constraints. Regarding the baseline GBDT LETOR LightGBM model, the focus has been on creating a feasible baseline model, not the best possible one. No hyperparameter tuning for the baseline GBDT LETOR model was performed. User-specific features (such as *is the brand of item A in user X's top 3 most bought brands?*) were not included, although it could have improved the results. The LightGBMRanker enables data grouping per user during the training in order to learn that the features have different importance for different users, but this would necessitate user-specific features in order to be able to exploit this during the inference. Only one type of negative sampling during the model training was used (random sampling), although other sampling techniques might have yielded better model results. Only one type of model was tested (LightGBMRanker). Due to time constraints and cost (potential revenue loss for Babyshop), the recommender systems were not deployed on any of the shopping platforms, and online evaluation through A/B testing or similar was not performed. Ideally, A/B testing a recommender system with weather features against the current production baseline should be done to evaluate the performance in the online setting.

Regarding the weather data, only historical weather data was used. Historical weather forecast data was available through the weather API used in this thesis, but the cost was deemed too high [17]. Using historical weather forecasts as features could possibly affect the result if forecasted weather has an impact on shopping behavior.

1.7 Ethics, Sustainability and Societal Aspects

Unnecessary consumption leads to higher use of the Earth's renewable and non-renewable resources and possibly other negative effects on the environment. This is not a unique problem to the fashion sector but concerns unsustainable diets, transportation, consumption of goods and services, etc.

Other environmental problems associated with fashion production and consumption include a high degree of materials being non-renewable, large carbon emissions and water needs, wastewater pollution, and the release of micro-plastics into the environment [18].

A set amount of different clothing is, of course, needed to meet the demands of the different seasons, but frivolous clothing consumption is common among individuals who are not poor. Effective recommender systems for children's fashion could lead to a higher degree of usage of bought products when the recommended products are highly relevant to the users. This could be positive for sustainability. Effective recommender systems could potentially decrease the number of returns, which has the potential to reduce emissions during shipping and promote a higher degree of material use.

Effective recommender systems could also lead to more clothing purchases per user, causing each bought product to have a lower degree of use. This is likely bad for the environment. To combat this, online retailers can remind consumers not to purchase products that they do not need or run programs that give incentives for recycling their purchased products.

When multiple products are displayed at once (e.g., when viewing products on the product listing pages) on an online shopping platform, the position is highly correlated to how likely the user is to click on a product, irrespective of how relevant the product actually is to the user. This is called *positional bias*. It has been shown through eye-tracking studies that users are less likely to look at items presented further down in vertical lists. A common user behavior is to click on the first product that seems relevant, before having looked through the full list of products [19]. Through the ordering of products on product listing pages or through other ways that products are ranked or recommended, users can be directed to a specific shopping behavior. This can be used in positive ways to present the most relevant products for the customers to support their decision-making by filtering out irrelevant products. This could also be used in questionable or malignant ways to direct the customers' shopping behavior toward items of interest by the e-tailer or a third party [20] (such as high-margin products, overstocked products or discontinued products). To act ethically, e-tailers should only serve users with product rankings and recommendations with benign intentions, that is, to only serve recommendations with the intent of fulfilling the user's true needs, wants and desires.

1.8 Structure of the Thesis

Chapter 2 presents relevant background information about the statistical methods used for the statistical analysis, contextual recommender systems, and the LightGBM library. Chapter 2 also summarises the related works reading in weather-sale data correlations, clothing recommendations based on weather & consumer psychology. Chapter 3 describes the methods and the data collection process. Chapter 4 describes the programs developed for the statistical analysis and recommender systems. Chapter 5 lists the results, which are commented in Chapter 6. Lastly, Chapter 7 lists the conclusions, limitations of this study, and potential future works.

Chapter 2

Background

This chapter provides basic background information about the statistical methods used in the analysis, contextual recommender systems, and how they can be evaluated. Additionally, this chapter summarises the related works reading in consumer psychology, weather-sale data correlations and clothing recommendations based on weather.

2.1 Statistical Analysis

This section will introduce the statistical methods used in this thesis. Statistical analysis was used to find correlations between weather and sales data, the directions, and strengths. These insights were used for implementing weather features in a recommender system. The statistical methods used are based on the methods from the related works (see Section 2.4.1) and include Pearson correlation coefficient, Spearman's rank correlation coefficient, and simple linear regression.

2.1.1 Pearson Correlation Coefficient

Pearson's product-moment correlation coefficient is a common measure for the degree of correlation between pairwise observations. The correlation coefficient provides the direction and strength of the linear relationship between two variables. It is based on the covariance but gives a normalized value and also provides the strength of the relationship. The value of the correlation coefficient ranges from -1 to 1 where -1 indicates a perfect negative linear relationship, 0 no linear relationship and 1 a perfect linear positive relationship [21, Chapter 2].

For two variables x and y, the Pearson correlation coefficient of a sample is formulated as:

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where $Cov(x,y)$ is the covariance, s_x and s_y are the sample standard deviations and n is the number of samples.

A relationship exists if [21, Chapter 2]:

$$|r_{xy}| \geqslant \frac{2}{\sqrt{n}}$$

2.1.2 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is another way to measure the degree of correlation between pairwise observations. It provides the strength of how well a monotonic function can describe a relationship between two variables and also the direction of the correlation. The coefficient is based on the ranks of each variable. Similarly to the Pearson correlation coefficient, the Spearman's rank correlation coefficient also ranges from -1 to 1 [21, Chapter 14], [22, Chapter 2].

For two variables x and y, the Spearman's rank correlation coefficient is formulated as:

$$r_s = \frac{Cov(R(x), R(y))}{s_{R(x)} s_{R(y)}}$$

where $Cov(R(x), R(y))$ is the covariance of the ranks of the variables, $s_{R(x)}$ and $s_{R(y)}$ are the sample standard deviations of the ranked variables.

2.1.3 Simple Linear Regression

Linear models can be used to model linear relationships or to approximate limited intervals of nonlinear relationships. The formula for a simple linear model is given by $Y = \beta_0 + \beta_1 X + \epsilon$, where Y is the dependent variable, X is the independent variable, β_0 and β_1 are the coefficients and ϵ is the random error term. In the context of this thesis, the difference in daily shares of sales of a Babyshop clothing category between two cities can be modeled as the dependent variable, and the difference in a weather variable (such as daily average temperature) between the same two cities can be modeled as the independent variable. Simple linear regression models are commonly developed using least square regression to find the best estimates of the coefficients β_0 and β_1 [21, Chapter 11].

Assumptions of simple linear regression include [21, Chapter 11]:

- The values of Y are linear function of X plus a random error term: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $(i = 1, ..., n)$
- The x_i values are real values, or if drawn from a random variable X, they are independent of the error terms so that for each i, $x_i(i = 1, ..., n)$ is independent of $\epsilon_i(i = 1, ..., n)$
- The error terms are random variables that have 0 mean and constant variance (homoscedasticity)
- The error terms are uncorrelated with each other.

To reject the hypothesis that the coefficient β_1 has no explanatory value, an F-test can be used. To reject the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$, the following F test statistic can be used:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

, where n is the sample size, $\hat{y}_i$ are the estimated y_i-values and $\bar{y}$ is the mean of the Y-variable. The decision rule for this test is: reject H_0 if $F \geq F_{1,n-2,\alpha}$ [21, Chapter 11].

2.2 Recommener Systems

This section gives a general overview of recommender systems and contextual recommender systems and describes how they can be evaluated.

2.2.1 Overview

Recommender systems rank items on product catalogs for users. They are important for aiding online shoppers in browsing the large quantities of products that are available in online stores. The low marginal cost of expanding product catalogs enables each online retailer to sell a diverse range of products, including niche products. The ability of online retailers to easily operate on a national or international level presents them to different customer groups with diverse demands, further complicating the tasks of effectively presenting relevant products to the customers [4, Chapter 5].

Recommender systems aim to present suitable items to users in environments when search intent is unclear or cannot easily be expressed. In some cases, product properties such as a person's clothing style or taste in music cannot easily be formalized. In other cases, users can be unaware of products and categories or not fully understand or remember their purchasing needs [4, Chapter 5].

Recommender systems aim to estimate the purchasing intent of the user based on indirect information such as purchase history, product popularity, and ratings. Based on information about products and users, a recommender system aims to calculate similarity metrics between products and users that are then used in creating recommendations. Three main types of similarity metrics are leveraged by recommender systems: user similarity, product similarity, and context similarity. User similarity aims to infer the purchasing intent of a user based on the historical behavior of similar customers. Product similarity aim to infer the purchasing intent of a user based on the attributes and categories that the customer has bought or browsed. Context similarity considers the user context such as time of day, date, location, season, etc. for inferring the purchasing intent. Information about catalog items can be used to calculate item similarities and information about the users can be used to calculate user similarities [4, Chapter 5].

The goal of recommender systems is to improve the user experience. This is done by producing a ranked list of items, from which a subset of top K ranked products or the full list can be shown to the users. A recommender system assumes the availability of customer ratings for all the items in the product catalog, either explicitly provided by their users or obtained through proxies such as purchasing and browsing histories. Several types of recommender systems exist, including content-based filtering, collaborative filtering, contextual, non-personalized, multi-objective, and hybrid recommender systems [4, Chapter 5].

2.2.2 Contextual Recommender Systems

Contextual recommender systems consider the circumstances of the user when the recommendations are made. Variables include user location, time, weather, demographical information, and promotion channel. Climate factors and local preferences can be accounted for using user location information.

For example, restaurant recommendations can present restaurants that are close to the user. A personal wardrobe recommendation can present suitable clothing based on the time of day (evening dresses will be recommended in the evening etc.) or weather (rainwear will be recommended when it rains etc.). The intent of the consumption can also be important: a person booking a hotel might have different preferences for vacation and business travel. A person going to the cinema might choose a different movie depending on if she goes by herself or together with her parents [23]. The use of contextual features allows users to be served with real-time recommendations based on their changing circumstances [4, Chapter 5].

2.2.3 Evaluation of Recommender Systems

Several objectives exist for recommender systems, these include [4, Chapter 5]:

- **Relevance:** Users should have a high inclination to interact with and purchase the recommended items
- **Novelty:** Recommender systems aim to show users available options. Non-trivial recommender systems should serve recommendations that are not already known to the users. A customer who has bought the first book in The Lord of the Rings trilogy is likely aware of the remaining books in the trilogy, but could derive value from being recommended other fantasy books
- **Serendipity:** Recommender systems should aim to help users find products that are novel and surprising and out of the categories of the items that the user typically buys.
- **Diversity:** A diverse set of recommended items can increase the chance that the user will find items of interest.

The focus objective of this thesis project is relevance and only the ranking relevancy will be evaluated. The other objectives require other metrics for evaluation.

Precision and recall-based metrics can be used to evaluate the relevancy of the top K recommendations. Precision measures the fraction of relevant recommendations of the recommended items and recall measures the fraction of relevant recommended items out of all the relevant items. It should be noted that the number of recommended items (K) affects the precision/recall tradeoff; a short list of recommended items is likely to leave out relevant items, while a long list is likely to have a large share of irrelevant items [4, Chapter 5].

Mean Average Precision at K (MAP@K) is a common metric for evaluating recommender systems and averages the $precision@k$ metric across all users for recommendation lists from length 1 to K and takes the arithmetic mean.

The formula is given as:

$$MAP@K = \frac{\sum_{u=1}^{U} AP@K(u)}{|U|} = \frac{\sum_{u=1}^{U} \frac{1}{|I_u|} \sum_{k=1}^{K} P@k(u) \cdot rel@k(u)}{|U|} = \frac{\sum_{u=1}^{U} \frac{1}{|I_u|} \sum_{k=1}^{K} \frac{Y_u(K) \cap I_u}{|Y_u(K)|} \cdot rel@k(u)}{|U|}$$

, where K is the number of recommended products, U is the set of users, $AP@K(u)$ is the average precision @ K for user u, $P@k(u)$ is the $precision@k$ for user u, I_u is the bought products buy user u in the test set, $Y_u(K)$ is the list of top K recommended items for user u. $rel@k(u)$ has the value 1 when the item at position k is relevant to user u and 0 otherwise [4, Chapter 5], [24], [25]

The Mean Reciprocal Rank (MRR) is the mean of the inverse of the ranks of the bought products. The MRR only considers the highest ranked item in the case that there are several correct items [26]. The formula is given as follows:

$$MRR = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{1}{rank_q}$$

, where Q is the set of queries evaluated.

2.3 Gradient Boosted Decision Trees and the LightGBMRanker Model

This section describes the LightGBM library and briefly explains key concepts about the training of the LightGBMRanker.

2.3.1 Learning to Rank

LETOR is the application of ML techniques in solving ranking problems as a form of supervised learning. Discriminative ML models designed for classification and regression can be adapted to the LETOR task [27]. Three main approaches of the LETOR task exist: the pointwise approach, the pairwise approach, and the listwise approach. The pointwise approach takes as inputs the feature vectors of each single item and outputs the degree of relevance for each item. The pairwise approach takes as input pairs of feature vectors of items and outputs the pairwise preference between each item pair. The listwise approach takes as input a set of items and outputs the ranked list of the items [28, Chapter 1]. The listwise approach is generally the most effective approach, followed by the pointwise approach [27]. These three approaches require different types of datasets, which typically limits which approaches that can be used for a specific problem.

Given a labeled training dataset, a learning-to-rank algorithm aims to find the optimal ranking model that minimizes the overall loss on the training dataset. For the learning process, either a loss function based on item relevancy labels and predicted scores can be used, or a function using ranking metrics as the learning objective can be used [27].

2.3.2 Gradient Boosted Decision Trees

Gradient boosted decision trees [29] is a model that can be used for the LETOR task (using regression trees). It consists of using gradient boosting on an ensemble of decision trees. It works by sequentially adding small, high-biased regression trees to the ensemble. Each additional tree aims to minimize the loss of the current ensemble [30], [31].

2.3.3 LightGBM

LightGBM is a library implementing gradient boosted decision trees developed by Microsoft Research [32]. It supports LETOR along with regression and classification. The main novel techniques introduced by the library are the Gradient-based One-Side Sampling and Exclusive Feature Bundling. The Gradient-based One-Side Sampling algorithm aims to speed up the training by performing a sampling to drop some of the training instances with small gradients, in order to focus more on instances with large gradients, which contributes more to the information gain. This is proven in the paper to outperform uniformly random sampling. At the same time, the algorithm aims to not change the original data distribution by much. The Exclusive Feature Bundling algorithm aims to speed up the training without degrading the accuracy by bundling sparse features (such as one-hot encodings). The optimal bundling problem is reduced to a graph coloring problem "(by taking features as vertices and adding edges for every two features if they are not mutually exclusive), and solving it by a greedy algorithm with a constant approximation ratio" [32]. LightGBM supports learning to rank using LambdaRank with the LightGBMRanker class.

2.3.4 LambdaRank

LambdaRank [33] [34] is a way to implement ranking evaluation metrics into the training in the pairwise LETOR approach. It was implemented to combat the mismatch between training and evaluation cost functions in ranking systems and allows a wide range of evaluation cost functions to be used during the training [33]. It enables cost functions that are non-differentiable or whose derivatives are flat everywhere (which is the case when sorting based on scores is introduced to the cost function). This is done by computing the gradients after the items have been sorted based on their relevancy scores. The main principle of LambdaRank is that during model training, the costs are not needed, only the gradient of the cost with respect to the relevancy scores [34]. The LambdaRank authors have shown empirically that LambdaRank optimizes the used evaluation metric [34]. LambdaRank is used in the LightGBM library for the LETOR task (when using LightGBMRanker) [35].

The loss function used in LambdaRank is defined as:

$$l(\mathbf{y}, \mathbf{s}) = \sum_{y_i > y_j} |\Delta Metric(i,j)|\, log_2(1 + e^{-\sigma(s_i - s_j)}) \tag{2.1}$$

Where $\mathbf{y}$ is the list of relevancy labels, $\mathbf{s}$ is the generated relevancy scores from the model, $|\Delta metric(i,j)|$ is the absolute difference of the training metric when items i and j are swapped and σ is a hyperparameter [27]. The training metric used in this study is Mean Average Precision (MAP).

2.4 Related Works

This section summarises the related works reading in weather-sales data correlation, clothing recommendations based on weather, and consumer psychology.

2.4.1 Weather Sales Data Correlation

Steinker et al. [36] studied the effect of local weather on aggregate e-commerce sales using Zalando sales data and weather data from the German state Hessen between October 2010 and May 2013. They found that sunshine, temperature and rain impacted daily sales, in particular "in the summer, on weekends, and on days with extreme weather" [36] and that this could be used to improve sales forecasting. The studied weather variables were total daily sunshine hours, daily average air temperature, and daily total precipitation. Their prestudy indicated that the inclusion of additional weather variables provided little marginal explanatory value for their model and that the other available weather variables (minimum and maximum air temperature, cloud cover, and barometric pressure, among others) were highly correlated with daily total sunshine hours, daily average air temperature, and daily total precipitation. Steinker et al. also studied the geographical granularity needed for the weather data. Six weather stations in different cities in Hessen were investigated (Bad Hersfeld, Frankfurt, Gießen, Kassel, Wiesbaden/Mainz, and Michelstadt), with an average distance between the stations of 49.9 km. By combining the three weather variables into a single index value and running a correlation analysis, they found that the weather at all six weather stations in Hessen correlated with each other all with Pearson's $r \geq 0.75$.

Badorf & Hoberg [37] studied the effect of weather on daily sales in a chain of 673 physical multi-category retail stores in Germany between 2013 and 2014. Their paper discusses how different weather parameters (temperature, precipitation, and sunshine) can be equally weighted to provide a single percentile value of good or bad weather based on the historical distributions of the weather parameters. Their paper mentions the seasonal effect of weather on sales (such as more sunlight having a different effect on sales in winter than in summer), and the result shows that weather seasonality is more important for explaining the variance than using a non-linear model to predict sales versus using a linear model. The study interestingly showed that good or bad weather had a different effect on sales depending on the season, with winter showing decreased sales for both bad and good weather, but the other seasons showing increased predicted sales in bad weather and decreased predicted sales in good weather. They also find that the effect of temperature, precipitation, and sunshine duration affect daily retail sales in a non-linear way by using non-linear models in their sales predictions. Results from the historical sales data showed that rain clothing and accessories sold well in abnormally bad weather (low temperature, low sunshine, and high precipitation) and badly in abnormally good weather (high temperature, a lot of sunshine, and no precipitation). The reverse was true for summer clothing.

Martínez-de-Albéniz & Belkaid [6] studied how the sales of coats and dresses varied with temperature using data from retail stores in 13 cities across four European countries between 2013 and 2014. The study finds that a 5 °C increase increases the estimated sales of dresses by 11 % and decreases the estimated sales of coats by 9 % using a log-linear model. Another interesting finding of this paper is that the weather does not seem to affect the customers' sensitivity to discounts.

Bahn & Kincade [7] studied the effect of weather on sales of women's branded business wear in 52 retail shops in the Seoul and Kyunggi areas in South Korea. They found that more seasonal garments were sold "during sales periods when drastic temperature changes occurred" [7].

Oh et al. [38] studied the correlation between Google Trends data for winter jackets and temperature. This way, they could study the time lag for winter jacket demand based on changing temperatures. The data was five years of Google Trends data (2014-2019) from the US state of New York and also air temperature data. The results suggested that search activities for winter jackets start after at least six continuous days of rapidly decreasing temperatures. A negative Person correlation was found between temperature and Google Trends searches for winter jackets ($r = -0.397, p = 0.01$). A linear regression model that fitted the winter jacket Google trends searches to the temperature also suggested the same thing.

Another study by Oh et al. [39] also studied correlations between Google Trends data for winter jackets and weather data. This time, 11 years of data was used (2008-2019). The wind chill effect showed the strongest correlation with Google Trends search activity for winter jackets using Pearson correlation ($r = -0.181, p = 0.000$), followed by maximum daily temperature ($r = -0, 169, p = 0.000$), average daily temperature ($r = -0.175, p = 0.000$), minimum daily temperature ($r = -0.176, p = 0.000$) and daily average wind speed ($r = 0.076, p = 0.005$). Using cross-correlation analysis, the authors found that the wind chill on the day before caused an increase in Google trends search activity the following day.

Back et al. [40] analyzed how sales of short-sleeved T-shirts and outerwear correlated with temperature in an online store in South Korea using sales data between 2014 and 2019. They found that sales of t-shirts increased with increasing temperatures and the sale of outerwear decreased with increased temperatures

Hong et al. [41] also analyzed an online store in South Korea for correlations between weather and sales using sales data from 2014 to 2018. They found a decrease in sales of winter items (scarves and socks) as temperatures increased and an increase in sales of short-sleeved T-shirts, shorts, and small bags as temperatures increased.

Kim et al. [42] studied the effect of weather factors on retail sales of the fall/winter line of a South Korean brand using data from 2012 to 2015. The interesting thing about this study is that it considers different types of brands: menswear brands, casual brands, and sports brands. Their statistical analysis includes correlation analysis, t-test, and multiple regression analysis. Results found include that outerwear (jackets, coats, and jumpers) sales increased with decreasing temperatures for all three kinds of brands. For pants, sales increased with decreasing temperature for menswear- and sportswear brands. For tops (shirts/blouses and T-shirts) and pants of menswear brands, a higher sea level pressure increased the sales volumes. An increase in humidity increased sales volumes of outerwear and tops for menswear and casual brands. A higher wind speed increased the sales volume of pants for casual brands.

Oh et al. [43] studied a national brand of casual clothing with shops all over South Korea and studied how the 7-day moving average temperatures correlated with 7-day moving average sales using data from 2013 to 2014. They found that “when 7-day moving average temperature value becomes 4 °C or higher, the growth period of S/S clothing sales starts. The peak period of S/S clothing sales starts at 17 °C, up to the highest temperature. When temperature drops below 21 °C after the peak temperature, the decline period of S/S clothing sales is over” [43]. The use of a 7-day moving average temperature enables the authors to capture synoptic scale meteorology phenomenons such as high pressure and low pressure, which typically have a time scale of one to three days [43]. The use of 7-day moving average sales removes weekday effects from the sales data (typically, more products are sold during the weekend).

Tian et al. [44] showed in their paper that weather conditions could be associated with the variety-seeking behavior of customers. They found that days with low sunlight, high temperature, and bad air quality induced more variety-seeking behavior among the customers. They also found that consumers with “fewer transaction records and longer inter-purchase times may be more likely to engage in variety seeking” [44]. The sales data came from supermarket panel data and transaction data from a large city in China between 2013 and 2014. The product categories considered studied juice, beer, yogurt, instant noodles, and biscuits.

The paper of Murray et al. [45] gave some interesting insights regarding sunlight and consumer spending. Using a mixture of methods and types of data, the effect of weather on consumer spending was analyzed. Using a lab study, the researchers modified the amount of sunlight that the test subjects were exposed to. Results indicate that sunlight decreases bad mood, which causes people to want to spend more money on the five examined categories green tea, orange juice, gym membership, airline ticket, and newspaper subscription. The study also shows that sunlight decreases negative affect.

In the review paper *Use of Weather Factors in Clothing Studies in Korea and its Implications: a Review* by Oh et al. [46], four research papers that could not be accessed through the internet or could not be machine translated into English were summarized. Using two years of daily sales data from a large discount store in South Korea, Hong et al. [47] found that sales of summer products increase with increasing temperature and sales of winter products increase with decreasing temperatures. They also found that the sales of swimsuits are related to precipitation and sales of winter mufflers are related to snowfall [47], [46]. Lee et al. studied sales data in six South Korean cities during one year for men's wear, women's wear, and casual wear. Among the three categories, they found that only men's wear was affected by the weather conditions (clear, cloudy, rainy, or snowy) [48], [46]. Jang & Lim [49] studied sales data for women's wear and men's wear in a department store in South Korea for three years, and the weather variables they investigated were temperature, precipitation, and average wind speed. They found that designer boutiques and unisex casual clothing sales were not affected by the weather [49], [46]. It needs to be noted that papers [43], [41], [40], [42] were read using Google Translate machine translation from South Korean to English, which is a source of error and misunderstanding.

2.4.2 Clothing Recommendations based on Weather

Liu et al. [50] built a system that can recommend clothes to users based on the weather and the user's personal wardrobe. It suggests using temperature, humidity, and wind scale for creating twelve distinct weather classes according to the weather-clothing classes suggested by China Meteorological Administration. The system can recommend the most suitable item from the user's submitted clothing items when weather information is given as input. The system can also, with a reference clothing item as input, find the best complementary items that also belong to the same weather category (i.e., lower-body clothing for an inputted upper-body clothing item). This paper is in the field of computer vision. Their method consists of classifying clothing attributes using a CNN and then mapping the clothing attributes to the twelve different weather classes. Clothing attribute labels were used as middle-level features to bridge the gap between low-level features of the clothes and high-level weather categories. The authors created the labeled Weather-to-Garment dataset containing the classes suits, upper and lower, with annotations describing the clothing items' attributes (e.g., color, material, and sleeve type) and also the suitable weather.

Wen et al. [51] constructed knowledge graphs of the user, clothing, and context for use in clothing recommendations. Weather information was considered for the context. The apriori algorithm was used to "capture the intrinsic correlations between clothing attributes and context attributes" [51]. Recommendations were generated by combining the user's requirements and the top-N algorithm. Their results showed that this approach could alleviate the cold start problem and perform well with regard to accuracy and precision in comparison with a collaborative filtering recommender system.

Yu-Chu et al. [52] uses a Bayesian network that recommends clothes based on (among other things) the weather context (season and temperature). Recommendations are made in the context of providing recommendations from a user's wardrobe. Based on the temperature and the season, different length of the sleeve is recommended.

Peifeng et al. [53] implemented an Android-based smart wardrobe system that has the ability to recommend weather-appropriate clothes from a person's wardrobe based on a person's geographical location and the weather for that location. Koshy et al. [54] developed a system to recommend clothes from a person's wardrobe based on skin complexion and season.

2.4.3 Consumer Psychology

The effect of weather on mood has been studied in several papers. The work of Keller et al. [55] indicates that comfortable weather increases people's mood.

The findings from field tests and laboratory studies from Zwebner et al. [56] suggest a temperature-premium effect across a variety of products. Their explanation is that "exposure to physical warmth activates the concept of emotional warmth, eliciting positive reactions and increasing product valuation" [56].

Bassi et al. [57] studied the effects of weather on financial markets and financial decision-making. Their findings suggest that sunshine and good weather lead to greater risk-taking behaviors in financial markets and that bad weather leads to greater risk aversion.

On the topic of purchasing in advance, Buchheim & Kolaska [58] studied the situation of advance purchases of outdoor movie tickets. They found that the weather at the time of purchase significantly influences the customer's purchasing decision, although the weather at the time of purchase is irrelevant to the cinema experience at a later date. This is interesting because it suggests that consumers can extrapolate weather expectations from irrelevant data or be exposed to projection bias or salience effects of the weather.

Chapter 3

Method

The purpose of this chapter is to provide an overview of the research method used in this thesis. Section 3.1 describes the research process. Section 3.2 details the data collection techniques used for this paper. Section 3.3 describes the experimental design. Section 3.4 explains the techniques used to evaluate the reliability and validity of the methods and the data. Finally, Section 3.5 describes the framework selected to evaluate the statistical analysis and the recommender systems.

3.1 Research Process

The research process used for this thesis was experimental research. Conducting experiments was deemed to be the most suitable way to provide an answer to the research question in section 1.2.1. By benchmarking the recommender system with weather features against the baseline without weather features, we could quantitatively measure the effect of the weather features on ranking performance.

3.2 Data Collection

3.2.1 Babyshop Data

3.2.1.1 Overview

Babyshop customer- and sales data for Stockholm, Göteborg, and Malmö regions were collected from the Babyshop data warehouse. This was then combined with the weather data from one location each in Stockholm, Göteborg, and Malmö. The reason for selecting data only from Sweden was partly to limit the scope of the analysis and partly to limit the effects of different sales campaigns in different countries. Within Sweden, all users shopping on the Babyshop platforms will see the same campaigns and discounts. This did not need to be accounted for by only selecting Swedish customers. Stockholm, Göteborg and Malmö areas were selected since these are the areas where Babyshop has the most customers. Selecting just a few cities meant that less weather data needed to be collected. These cities were also selected since they are not close to each other, which means that the local weather was likely to differ.

A data point (a single customer and a single item) fed into the ranking model contained the the city of the customer, static features about the item, and variable features of the item based on the order date. In evaluation, bought items were grouped together per user and day to form a shopping basket when evaluating the MAP@K. This assumes that all products bought by a user on a day belong to the same purchase. However, some users make multiple purchases on the same day. The product data was filtered to only include clothing, bags, and shoe products. Data from 2021-01-01 to 2022-12-31 was used in the recommender systems. The sales data was filtered so that only customers with at least two orders were included.

3.2.1.2 Features

The Babyshop Snowflake data warehouse contains data about the customers and the products that were used for this thesis. Some of the features were already available in the existing tables in the data warehouse, and others had to be computed (using SQL). The only user information used was the city of the user, the customer number, and the date of bought products. The users cannot be uniquely identified from this information, and all personally identifiable information was hashed.

The product data contains information about the products and their categories. Based on historical sales figures on the Babyshop platforms, sales trend features for each item were calculated. This was done on the Babyshop Snowflake SQL platform to make it feasible to compute in a reasonable time due to the high scalability of Snowflake using parallel query execution. The sales trends features were calculated using sales data from all of Sweden. The customer and product features are described in Table 3.1 and the calculated product features in Table 3.2.

Negative sampling was used during the training. For every bought item per customer, an item that was not bought by the customer on the same day was sampled and given the label 0. The sampling method that was performed was uniform random sampling. This was chosen since it was easy to implement in SQL.

Feature Name	Description	Examples	Type
CUSTOMER_NO	Customer number		Categorical
ORDER_DATE			datetime64 [59]
CITY	Customer postal town/city	Stockholm, Malmö, Göteborg	Categorical
BEX_PRODUCT_NO	Product number		Categorical
LABEL	1 if a product was bought by the customer, 0 if the product was negatively sampled		Binary
COLOUR		red,yellow, beige	Categorical
GENDER_DESCRIPTION		boys, girls, unisex	Categorical
BRAND		Kuling, Gap	Categorical
COLLECTION		autumn/winter 2021, spring/summer 2021	Categorical
PRODUCT_GROUP_LEVEL_1_DESCRIPTION	Highest level product description	clothing, accessories, footwear	Categorical
PRODUCT_GROUP_LEVEL_2_DESCRIPTION	More detailed product description	scarves, tops, shoes, underwear, bottoms	Categorical
PRODUCT_GROUP_LEVEL_3_DESCRIPTION	Most detailed product description	jeans, wellingtons, t-shirts, fleece jackets	Categorical

Table 3.1: Sales data features

Feature Name	Description	Type
1D_SALES_QTY_PRODUCT	Number of sales of the product the day before	Numerical
14D_SALES_QTY_PRODUCT	Number of sales of the product from 1 day before to 14 days before	Numerical
14D_SALES_CHANGE_PRODUCT	Percentual change between the sales of the product between 1-14 days before and 15-28 days before	Numerical
1D_SALES_QTY_BRAND	Number of sales of the brand the day before	Numerical
14D_SALES_QTY_BRAND	Number of sales of the brand between 1 and 14 days before	Numerical
14D_SALES_CHANGE_BRAND	Percentual change between the sales of the brand 1-14 days before and 15-28 days before	Numerical
1D_SALES_QTY_CAT3	Number of sales of the PRODUCT_GROUP_LEVEL_3_DESCRIPTION the day before	Numerical
14D_SALES_QTY_CAT3	Number of sales of the PRODUCT_GROUP_LEVEL_3_DESCRIPTION between 1 and 14 days before	Numerical
14D_SALES_CHANGE_CAT3	Percentual change between the sales of the PRODUCT_GROUP_LEVEL_3_DESCRIPTION 1-14 days before and 15-28 days before	Numerical
14D_TOP_100_HIT	1 if the product is in the top 100 sold items in the period 1 to 14 days before, else 0	Binary
14D_TOP_10_HIT	1 if the product is in the top 10 sold items in the period 1 to 14 days before, else 0	Binary

Table 3.2: Sales data calculated features

3.2.1.3 Location Based Data Filtering

The customer data in the Babyshop data warehouse lists the users' town/city based on the postal town system used in Sweden (postort in Swedish). For the small dataset, only users with the postal towns Stockholm, Göteborg and Malmö were selected.

For the large dataset, data from customers from a larger geographical area was included. Based on the findings of Steinker et al. as presented in section 2.4.1, postal towns from a larger geographical area were collected. Since Steinker et al. found that weather observational stations with an average distance between them of 49.9 km had highly correlated weather, a similar distance was used for the user data selection for the Stockholm, Göteborg, and Malmö areas. The method used was to use the website MAPS.ie [60] to draw circles of radius 25 km from the centers of Stockholm, Göteborg, and Malmö (by using the default centers used by the MAPS.ie website, see Appendix B for coordinates). The municipalities fully covered or mostly covered by the circle were included in the selection. From the names of the municipalities, the corresponding postal towns were scraped and processed from the website post24.se [61]. The full list of municipalities and postal towns used is listed in

Appendix A. The reasoning behind selecting 25 km as the radius of the circles was to ensure that users were not much more than 50 km apart. This was based on the findings of Steinker et al. This method was not very precise since municipalities mostly but not fully covered by the circles were included. This means that the upper limit of the potential distance between customers in the same city has an upper bound above 50 km. The number of bought products for the small and the big datasets are listed in Table 3.3.

Dataset	Number of orders
Small Dataset	390 000
Big Dataset	967 000

Table 3.3: Number of orders per dataset

3.2.1.4 Sales Data for Statistical Analysis

For the statistical analysis, sales data from a longer time period could be used since the required sales data was already available in the Babyshop data warehouse. Data from 2018-01-01 to 2023-05-21 was included, but not all days were always used in the statistical analysis since sales in the examined clothing category are required in both Stockholm and Göteborg for a day to be included. The sales data lists the daily quantity sold per *PRODUCT_GROUP_LEVEL_3_DESCRIPTION* (the most detailed product description, see Table 3.1) for Stockholm and Göteborg areas. The included postal towns were the same as described in Section 3.2.1.3. The total number of sold products across all categories per day was also collected in order to be able to calculate the share of sold products that belong to the clothing categories of interest.

3.2.2 Weather Data

Observational weather data was used in this study. The data was obtained using the Visual Crossing API [62]. The default weather station selection was used, and Stockholm, Göteborg, and Malmö were entered as the locations. Weather data was collected for the same time period as the sales data. Weather data from several weather stations and station types are aggregated using the Visual Crossing API [63]. A weighted average based on several factors is then used to calculate the values that are provided by the API. Data from surface observational stations are weighted more heavily than satellite or remote data in the API. The weather stations are also weighted based on their distance to the location and the overall quality of the weather station [63]. The reasons for choosing this API were that it was simple to use and it had all the desired weather variables. The weather variables used in the thesis are listed in Table 3.4, and their descriptions are listed on the Visual Crossing website [64].

Weather variable	**Description**	**Unit**
feelslike	Daily average feels-like-temperature. The feels-like temperature is a combination of the heat index at high temperatures, the temperature at medium temperatures, and the wind chill index at low temperatures	C
precip	Daily total precipitation	mm
solarenergy	The total daily energy from the sun	MJ/m^2
windspeed	Daily maximum wind speed	km/h
snow	The daily amount of snow precipitation	cm
snowdepth	Average daily snow depth	cm

Table 3.4: Weather features

3.2.3 Date Features

Date features were included in order to try to capture seasonal trends, preferences, and user engagement patterns. The features are listed in Table 3.5.

Feature
dayofyear
dayofweek
month
week

Table 3.5: Date Features

3.3 Experimental Design

3.3.1 Test Environment

The programs for the statistical analysis and the recommender systems were written in Python 3.10. Sales trend features were calculated in Snowflake SQL in the Babyshop data warehouse. The programs used for the statistical analysis and the recommender systems are provided in the thesis GitHub repository [65]. The programs will be described in Chapter 4.

3.3.2 Software and Hardware Environments

The full Anaconda virtual environment is listed in the thesis Github repository [65]. The main Python libraries used are LightGBM, pandas, scikit-learn, statsmodels and SciPy along with plotting libraries (plotly, matplotlib, and seaborn). The libraries are described in detail below:

- **LightGBM**: GBDT-library used for creating the ML ranking models
- **matplotlib and seaborn**: Used for plotting
- **pandas**: Data Analysis and data manipulation library
- **plotly**: Used for plotting and creating linear regressions
- **scikit-learn**: Used for testing assumptions linear regression and for splitting datasets
- **SciPy**: Used for statistics
- **statsmodels**: Used for testing assumptions of linear regression

All programs and configurations except for using the large dataset in the recommender systems program were run on a quad-core laptop with 16 GB RAM. In order to run the program for the recommender systems using the large dataset, more RAM was required. For this, Google Cloud Platform was used, with a virtual machine with 128 GB RAM memory and 16 CPU cores.

3.4 Assessing Reliability and Validity of the Methods and the Data

3.4.1 Validity of Method

3.4.1.1 Statistical Analysis

The statistical analysis to find potential correlations between local weather and sales on the Babyshop platforms could potentially have been conducted in many different ways. Exploratory data analysis was first performed to analyze and visualize the data in different ways. Other methods of performing the analysis were tried, but they were deemed too noisy to be able to extract any meaningful results. The chosen method used data from Stockholm and Göteborg and processed the weather and sales data into the two following variables:

- **X:** Difference in the daily value for a weather variable (such as wind-speed or snowfall) in Stockholm - Göteborg ($weather\_variable_{day_d,Stockholm} - weather\_variable_{day_d,Göteborg}$)

- **Y:** Percentual increase between the daily share of sold items of a clothing category in Stockholm compared to the daily share of sold items of the same clothing category in Göteborg (using t-shirts as the category, a value of 100 would mean that the daily share of t-shirts out of total sales in Stockholm is 100% bigger than the share in Göteborg)

From this, Pearson correlation coefficients and Spearman's rank correlation coefficients, and linear regression between the two variables could be calculated.

This method was able to generate some meaningful correlations, but it has some drawbacks. This method essentially only tells how the difference in weather between Stockholm and Göteborg correlates with the difference in the share of sold items in Stockholm and Göteborg. Differences in baseline preferences for different clothing items between Stockholm and Göteborg were not accounted for. Running the statistical analysis for multiple cities would have been helpful for increasing the validity of the statistical analysis.

The intent of using this method was to show how differences in weather correlate with differences in sales. By comparing weather and sales in two cities on the same day, this method aimed to show how the sale differs when the weather differs.

3.4.1.2 Recommender Systems

Examining the training and evaluation performance during the training of the LightGBMRanker models gave us confidence that they were learning as they were supposed to. By comparing the evaluation performance of the LightGBM rankers against the naive baselines (a random ranker and a ranking based on the popularity of the bought products), it was possible to evaluate the ranking performance.

The study aimed to measure the quality of product rankings. The rankings were evaluated using standard metrics for evaluating ranking relevancy. On the overall evaluation of a recommender system, there are also other qualities of recommender systems besides the ranking relevancy that were not evaluated as a part of this thesis (see Section 2.2.3).

3.4.2 Reliability of Method

3.4.2.1 Statistical Analysis

Only data from Stockholm and Göteborg was compared. This means that the same correlation might not be found between other cities. The findings might not generalize to all of Sweden. It would have been better to test for correlation between several cities rather than just two.

3.4.2.2 Recommender Systems

Customer and weather data from only Stockholm, Göteborg, and Malmö were used. The findings might not generalize to all of Sweden.

3.4.3 Reliability of Data

It could have been possible to use more than 365 days of training data for the recommender systems. The reason for selecting 365 days of data was that changes in consumer trends and tastes change what products are bought over time. Also, not all items are restocked by Babyshop year after year. This makes the usefulness of training the recommender systems on old data questionable. Using 365 splits for evaluation means that the evaluation day is averaged across all varying weather conditions within a year.

3.5 Evaluation Framework

3.5.1 Statistical Analysis

Linear regressions, Pearson correlation coefficients, and Spearman correlation coefficients were used to evaluate what weather variables correlated with sales on the Babyshop platforms. An overall judgment based on all three factors decided which variables were adopted as features for the recommender system. We also compared findings from other studies in weather and sales data correlation (see Section 2.4.1) against the findings using this method and this data.

3.5.2 Recommender Systems

Evaluation based on the selected evaluation metrics: MRR and MAP@10 against the ML-baseline and the naive baselines showed if including weather features improved the ranking relevancy.

Chapter 4

Implementation

4.1 Statistical Analysis

The program was developed as a script that takes as input a clothing category (*PRODUCT_GROUP_LEVEL_3_DESCRIPTION*) from the Babyshop data and also a weather variable. This way, different configurations could easily be tested.

The weather and sales data were loaded into pandas DataFrames. The shares of sold items were calculated using the daily counts of items sold from the specified category and the daily counts of total sales from all item categories. The dates were filtered so that only dates that had sales of the clothing category in both Stockholm and Göteborg were included.

The program stored these variables in a DataFrame:

- **X:** Difference in the daily value for the weather variable in Stockholm - Göteborg ($weather\_variable_{day_d,Stockholm} - weather\_variable_{day_d,Göteborg}$)
- **Y:** Percentual increase between the daily share of sold items of a clothing category in Stockholm compared to the daily share of sold items of the same clothing category in Göteborg (using t-shirts as the category, a value of 100 would mean that the daily share of t-shirts out of total sales in Stockholm is 100% bigger than the share in Göteborg)

At this point, there was one data point per day. The data points were duplicated based on the total number of sales per day in Stockholm and Göteborg in order to linearly scale the weight of each data point with its sample size. This was done so that the larger variances from days with smaller sample sizes (fewer items sold) have less effect on the result of the statistical analysis.

The Y-values were clipped in the (-100,100)-range to filter out extreme results. For snow, snowdepth, and precipitation, the data points where the difference between Stockholm and Göteborg was zero were removed. This was done because there is no rain or snow on most days.

The program calculated Pearson correlation, Spearman correlation, and linear regression. It also did several graphical tests for the assumptions of linear regression. The code for the graphical tests for the assumptions of linear regression was taken from a Kaggle notebook developed by Shruti Iyyer [66].

The program also considered the same data for simple moving averages of 2,3,5,7,10, and 14 days of both variables (abbreviated as SMA_2, SMA_3...). For each Simple Moving Average (SMA), the main loop of the program was repeated, and the same outputs were generated. The simple moving averages calculations are displayed in Appendix C. The program was only evaluated using the most popular clothing categories. This was done so that there were enough data points per category.

4.2 Recommender Systems

4.2.1 Parameters

The program was developed as a script that takes several parameters: number of splits, number of days per training split, and start date. A Boolean parameter toggled the program to include the weather data.

4.2.2 Data Preprosessing

The pandas.DataFrame holding the sales data has date features calculated from the order dates. The sales count features were normalized based on the highest value per day. This was done because daily sale counts vary greatly. For example, days with large discounts typically have significantly larger sales than the average day.

A function splitted the sale data into a list of dictionaries containing the training and evaluation data for each split. This function also encoded the categorical columns. K different values in a categorical column were encoded using integers 0 to $K-1$. This is preferred over one-hot-encoding in LightGBM.LGBMRanker [14].

4.2.3 Features

Based on the findings from the statistical analysis (see section 5.1), feelslike, precip, solarenergy, windspeed, snow, snowdepth were selected as weather features along with their simple moving averages for 3 and 7 days, the differences between the daily weather and the weather of the day before, and also the difference between the daily weather and the 7-day simple moving average from the day before. With windspeed as an example, the calculated weather features are described in Table 4.1.

Weather Feature Name	**Description**
windspeed	Daily maximum wind speed
windspeed_1D_difference	Difference between daily average wind speed the current day and the day before
windspeed_SMA3	3 days simple moving average of daily average wind speed
windspeed_SMA7	7 days simple moving average of daily average wind speed
windspeed_SMA7_difference	Difference between daily average wind speed and the 7 days simple moving average for the preceding 7 days

Table 4.1: Descriptions of weather features using windspeed as the example feature

The used features for the LightGBMRanker model with weather features are displayed in Table 4.2 (with descriptions given in Tables 3.1, 3.2 & 4.1). The baseline LightGBMRanker model uses the same features except for the weather features. Figure 4.1 shows the feature correlation matrix.

Feature
CITY
COLOUR
GENDER_DESCRIPTION
BRAND
COLLECTION
PRODUCT_GROUP_LEVEL_1_DESCRIPTION
PRODUCT_GROUP_LEVEL_2_DESCRIPTION
PRODUCT_GROUP_LEVEL_3_DESCRIPTION
1D_SALES_QTY_PRODUCT
14D_SALES_QTY_PRODUCT
14D_SALES_CHANGE_PRODUCT
1D_SALES_QTY_BRAND
14D_SALES_QTY_BRAND
14D_SALES_CHANGE_BRAND
1D_SALES_QTY_CAT3
14D_SALES_QTY_CAT3
14D_SALES_CHANGE_CAT3
14D_TOP_100_HIT
14D_TOP_10_HIT
dayofyear
dayofweek
month
week
feelslike
feelslike_1D_difference
feelslike_SMA3
feelslike_SMA7
feelslike_SMA7_difference
precip
precip_1D_difference
precip_SMA3
precip_SMA7
precip_SMA7_difference
snow
snow_1D_difference
snow_SMA3
snow_SMA7
snow_SMA7_difference
snowdepth
snowdepth_1D_difference
snowdepth_SMA3
snowdepth_SMA7
snowdepth_SMA7_difference
solarenergy
solarenergy_1D_difference
solarenergy_SMA3
solarenergy_SMA7
solarenergy_SMA7_difference
windspeed
windspeed_1D_difference
windspeed_SMA3
windspeed_SMA7
windspeed_SMA7_difference

Table 4.2: Features used for the LightGBM + weather features model

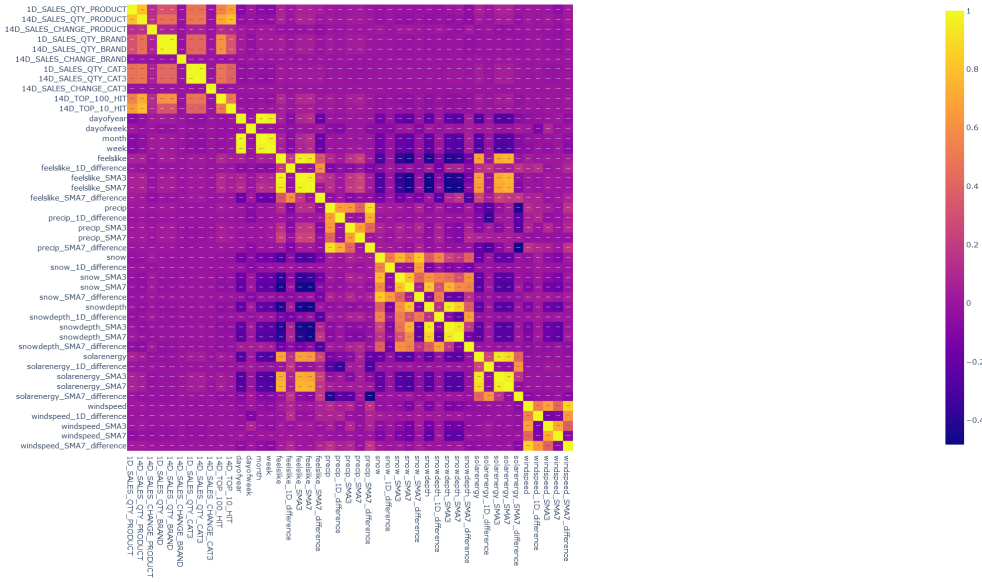

Figure 4.1: Feature correlation matrix (first data split, large dataset)

4.2.4 LightGBM.LGBMRanker Training

For each split, a separate ranker was trained, and the intermediate results were saved. The training window consisted of 365 consecutive days, and the evaluation was performed on the 366th day. 365 data splits were used, with the first split starting the training window on 2021-01-01 and the 365 starting the training window on 2021-12-31. The ranker parameters are displayed in listing 4.1:

```
ranker = LGBMRanker(
            objective="lambdarank",
            metric="map",
            boosting_type="gbdt",
            max_depth=7,
            n_estimators=300,
            importance_type="gain",
            verbose=1000,
        )
```

Listing 4.1: LightGBM.LGBMRanker parameters

These parameters were taken from Alexander Vishnevskiy's submission to the H&M Personalized Fashion Recommendations competition [9]. This was deemed suitable as both are similar problems. The metric parameter was changed to MAP as it was deemed more suitable to this problem since the data has no graded relevancy.

The grouping of the data and the LightGBM.LGBMRanker.fit() parameters are displayed in listing 4.2:

```

X_train_groups = (
            X_train.copy()
            .groupby(["ORDER_DATE", "CUSTOMER_NO"])["
    BEX_PRODUCT_NO"]
            .count()
            .values
        )

ranker = ranker.fit(
            X=X_train,
            y=y_train,
            group=X_train_groups,
            feature_name=X_train.columns.tolist(),
            categorical_feature=[
                "CITY",
                "COLOUR",
                "GENDER_DESCRIPTION",
                "BRAND",
                "COLLECTION",
                "PRODUCT_GROUP_LEVEL_1_DESCRIPTION",
                "PRODUCT_GROUP_LEVEL_2_DESCRIPTION",
                "PRODUCT_GROUP_LEVEL_3_DESCRIPTION",
            ],
            eval_set=[(X_train, y_train), (X_test, y_test)],
            eval_metric=["binary_error", "map"],
            eval_group=[X_train_groups, X_test_groups],
            callbacks=[early_stopping(stopping_rounds=50,
    first_metric_only=True)]
        )
```

Listing 4.2: LightGBM.LGBMRanker.fit() parameters

4.2.5 Naive Baselines

The popular ranker ranked products according to their popularity in the last 30 days of the training data. The random ranker produced a random permutation of the ranking list produced by the popular ranker.

4.2.6 Inference

The data used for inference followed the same preprocessing steps as the data used for the training. The evaluation was performed on different levels. MRR was calculated for each item for each user. This is not the standard way to use it, but we wanted to evaluate each item, not just the most highly-ranked item that the customer bought. The MAP@K for K = 10 was calculated for each customer and day. The results were averaged across all splits in the end using the arithmetic means. The Python code for calculating MAP@K was developed by Ben Hamner [67]. The number of products that were considered for the LightGBM ranking was limited in order to speed up the program and decrease RAM usage. The product selection was the 1000 most popular products in the training data per split.

Chapter 5

Results

5.1 Statistical Analysis

Table 5.1 shows correlations from the related works in weather and sales correlations (section 2.4.1) that could be retested on the Babyshop data and using our method for statistical analysis. Some of the findings could be reproduced.

Study	Found Correlation	Result On Babyshop Data
Martínez-de-Albéniz & Belkaid [6]	Sales of dresses increase with increasing temperatures.	Sales of casual dresses show no clear correlation with feels-like temperature.
Back et al. [40]	Sales of t-shirts increase with increasing temperatures.	There is no clear trend for how t-shirts correlate with feels-like temperature.
Back et al. [40]	Sales of outerwear increase with decreasing temperatures.	Padded and puffer jackets sell as a smaller share in Stockholm when it is warmer in Stockholm. Shell pants sell a larger share in Stockholm when it is warmer in Stockholm
Hong et al. [47]	Shorts sales increase as temperature increases.	Short sales sell a larger share in Stockholm when it is warmer in Stockholm.
Hong et al. [47]	Sales of summer products increase with temperature.	Sun hats sell a larger share in Stockholm when Stockholm is warmer. Sunsuits sell a larger share in Stockholm when Stockholm is warmer.
Hong et al. [47]	Sales of winter products decrease with increasing temperatures.	A smaller share of snow boots is sold in Stockholm than in Göteborg when Stockholm is warmer. A smaller share of ski gloves and mittens is sold in Stockholm than in Göteborg when Stockholm is warmer
Hong et al. [47]	Sales of swimsuits are related to precipitation.	Swimsuits show no clear correlation with precipitation.

Table 5.1: Findings from related works retested using the Babyshop data and the method of this thesis

The tested configurations in the statistical analysis and the interpreted correlations are listed in Table 5.2. A positive correlation means that the share of sold items of the category is larger in Stockholm than in Göteborg when the difference in the weather variable is positive. A negative correlation means that the share of sold items of the category is larger in Göteborg than in Stockholm when the difference in the weather variable is positive. Examples of supporting material for the statistical analysis results are listed in Appendix D. Some of the found correlations are in the expected direction, but some surprising results are found, for example, that ski gloves and mittens sell as a larger share in Stockholm when the feelslike temperature is higher in Stockholm.

PRODUCT_GROUP_3_DESCRIPTION	Weather Variable	Direction and strength of correlation
Beanies	Feelslike	No meaningfull correlation
Casual dresses	Feelslike	No meaningfull correlation
Casual dresses	Solar energy	Negative correlation
Casual trainers	Precip	Negative correlation
Fleece jackets	Feelslike	Negative correlation
Fleece jackets	Windspeed	Negative correlation
Rain sets	Feelslike	Stable negative correlation
Rain sets	Precip	Stable positive correlation
Rain sets	Solar energy	Stable negative correlation
Shell jackets	Feelslike	Positive correlation
Shell jackets	Precip	Positive correlation
Shell jackets	Windspeed	Positive correlation
Shell pants	Precip	No meaningfull correlation
Shell pants	Snow	No meaningfull correlation
Shell pants	Snow depth	Positive correlation
Shell pants	Windspeed	Stable positive correlation
Shorts	Feelslike	Positive correlation
Shorts	Precip	No meaningfull correlation
Shorts	Solar energy	No meaningfull correlation
Ski gloves and mittens	Feelslike	Positive correlation
Ski gloves and mittens	Snow	Stable positive correlation
Ski gloves and mittens	Snowdepth	Stable positive correlation
Snow boots	Feelslike	Stable negative correlation
Snow boots	Snow	Stable positive correlation
Snow boots	Snow Depth	Stable positive correlation
Snow boots	Solar energy	No meaningful correlation
Strappy sandals	Feelslike	Stable positive correlation
Strappy sandals	Solar energy	Stable positive correlation
Sun hats	Feelslike	Stable positive correlation
Sun hats	Solar energy	Stable positive correlation
T-Shirts	Feelslike	No meaningful correlation
Welingtons	Feelslike	No meaningfull correlation
Wellingtons	Precip	Stable positive correlation
Wellingtons	Solarenergy	Stable negative correlation
Winter coveralls	Feelslike	Negative correlation
Winter coveralls	Snow	Stable positive correlation
Winter coveralls	Snowdepth	Positive correlation

Table 5.2: Summarised results of the statistical analysis

Using the results from the daily values and simple moving averages for 2,3,5,7,10 and 14 days, the averaged results and averaged p-values per configuration are reported in Table 5.3. The F-test p-value column depicts the probability that the coefficient in the linear regression is equal to 0 (see section 2.1.3) and is a way to imply the meaningfulness of the regression (a lower value is better). The highest absolute value of the average Pearson correlation coefficient is 0.31 (sun hats & feelslike), and for the average Spearman correlation coefficient this value is 0.33 (strappy sandals & feelslike). The Pearson and Spearman correlation coefficients are generally of the same magnitude and direction per configuration. Most configurations have low enough F-test p-values to imply that the linear regressions are meaningful.

PRODUCT Description	Weather Variable	Pearson correlation (p-value)	Spearman correlation (p-value)	Linear regression coefficient	Linear regression constant	F-test p-value
Beanies	Feelslike	0.05924 (0.00274)	0.06051 (0.04196)	1.40382	26.18451	0.00275
Casual dresses	Feelslike	-0.01047 (0.147782)	-0.00164 (0.17744)	-0.24028	28.88794	0.14778
Casual dresses	Solar energy	-0.13503 (0.10023)	-0.15209 (0.03951)	-2.54313	28.35512	0.10018
Casual trainers	Precip	-0.03327 (0.04090)	-0.07872 (3.13285e-9)	-0.38486	18.18854	0.04086
Fleece jackets	Feelslike	-0.06995 (0.12499)	-0.06295 (0.07714)	-1.41521	7.52735	0.12498
Fleece jackets	Windspeed	-0.10596 (4.00155e-6)	-0.10627 (2.45328e-6)	-1.04042	10.76240	4.00052e-6
Rain sets	Feelslike	-0.17138 (1.87517e-103)	-0.19869 (3.98494e-116)	-3.67255	0.96130	1.87142e-103
Rain sets	Precip	0.18834 (7.03820e-209)	0.23471 (0.0)	2.21763	8.80153	7.04285e-209
Rain sets	Solar energy	-0.21976 (0.0)	-0.24352 (0.0)	-3.51057	2.57599	0.0
Shell jackets	Feelslike	0.07757 (0.0)	0.07377 (0.00027)	1.51285	-0.88342	0.0
Shell jackets	Precip	0.03519 (0.27798)	0.05926 (0.00031)	0.43301	-2.40981	0.27808
Shell jackets	Windspeed	0.07678 (0.23539)	0.10865 (0.0)	0.78512	-4.72402	0.23546
Shell pants	Precip	0.00687 (0.01857)	-0.01799 (0.05785)	0.1535	12.94433	0.01862
Shell pants	Snow	-0.05157 (0.24507)	-0.05823 (0.15603)	-5.67914	0.91409	0.24509
Shell pants	Snow depth	0.15286 (0.00018)	-0.04245 (0.06146)	3.18464	-7.37838	0.00018
Shell pants	Windspeed	0.15001 (0.0)	0.17988 (0.0)	1.7507	9.75311	0.0
Shorts	Feelslike	0.24953 (0.0)	0.20376 (1e-05)	5.36016	54.86345	0.0
Shorts	Precip	-0.05224 (0.19392)	-0.08167 (0.11727)	-1.65082	49.84167	0.19394
Shorts	Solar energy	0.01023 (0.04569)	0.06149 (0.22011)	0.07929	50.68712	0.04566
Ski gloves and mittens	Feelslike	-0.15923 (0.0)	-0.15518 (0.0)	-4.48774	22.3666	0.0
Ski gloves and mittens	Snow	0.21492 (5e-05)	0.18671 (0.0)	8.09023	30.41464	5e-05
Ski gloves and mittens	Snowdepth	0.16535 (1e-05)	0.25489 (0.0)	1.94542	30.71275	1e-05
Snow boots	Feelslike	-0.18432 (0.0)	-0.19916 (0.0)	-3.74381	5.95394	0.0
Snow boots	Snow	0.18811 (0.0)	0.1709 (0.03993)	6.15737	19.4702	0.0
Snow boots	Snow depth	0.07783 (0.10158)	0.1353 (0.00347)	0.68289	20.99572	0.10157
Snow boots	Solar energy	0.03336 (0.00172)	0.09374 (0.0592)	0.78258	14.76718	0.00171
Strappy sandals	Feelslike	0.26347 (0.0)	0.33254 (0.0)	5.03835	16.54128	0.0
Strappy sandals	Solar energy	0.14959 (0.0)	0.13038 (0.0)	1.38787	13.46281	0.0
Sun hats	Feelslike	0.30952 (0.0)	0.31513 (0.0)	6.54958	21.93048	0.0
Sun hats	Solar energy	0.25019 (0.0)	0.25501 (0.0)	2.68119	16.79583	0.0
T-shirts	Feelslike	-0.00924 (0.00474)	0.01064 (0.20581)	-0.2171	30.32601	0.00473
Welingtons	Feelslike	0.02061 (0.15773)	0.02628 (0.14431)	0.42919	6.98075	0.1577
Wellingtons	Precip	0.27035 (0.0)	0.29833 (0.0)	3.17638	10.62193	0.0
Wellingtons	Solarenergy	-0.23989 (0.0)	-0.24972 (0.0)	-3.24314	2.52166	0.0
Winter coveralls	Feelslike	-0.03006 (0.03735)	-0.01283 (0.13546)	-0.54077	11.07237	0.03729
Winter coveralls	Snow	0.19108 (0.0)	0.27423 (0.0)	5.9097	12.89862	0.0
Winter coveralls	Snowdepth	0.14443 (0.0)	0.20935 (0.0)	1.07695	13.25877	0.0

Table 5.3: Average results per configuration of the statistical analysis

A few examples of linear regression for different configurations for daily and simple moving average values are displayed in Figure 5.1. A bit of a trend in the direction of the line can be seen on some plots, but the variance is typically quite high. This is also in line with the low degrees of Pearson and Spearman correlations as listed in Table 5.3.

Figure 5.1: Examples of linear regressions for daily values and simple moving averages

Based on the result of the statistical analysis, all examined weather variables were featurized for the recommender system (see Table 5.4). In addition to the daily values of these variables, we generated several other features for each one. We decided to add the three- and seven-day moving averages after reviewing the results for moving averages of the sales and weather data. We also incorporated the differences between the daily weather and the weather of the day before, as well as the difference between the daily weather and the seven-day simple moving average of the seven preceding days. These features were added since we believed that they could have predictive power. The weather features are described in detail in Table 4.1

5.2 Recommender Systems

The features used in the LightGBM Ranker model and their importance are displayed in Figure 5.2. The values displayed are the total information gain of the tree splits that use that feature [35]. The highest-ranked weather feature is at rank 12 (solarenergy_SMA7).

Weather variable
feelslike
precip
solarenergy
windspeed
snow
snowdepth

Table 5.4: Weather variables featurized for the LightGBM recommender system

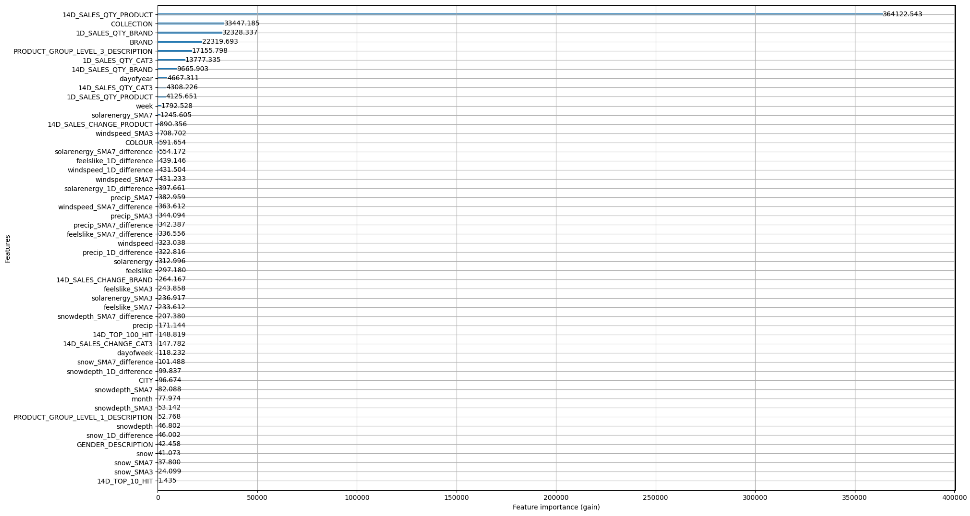

Figure 5.2: Feature importance ranking (first split, training on the large dataset)

The ranking performances on the small and the large datasets are displayed in Tables 5.5 and 5.6. It can be noted that the LGBMRanker + weather features always beats the baseline LGBMRanker model. On the small dataset, the increase is 1.26% for MRR and 2.23% for MAP@10. On the large dataset, the increase is 1.34% for MRR and and 1.11% for MAP@10. The LightGBM rankers beat the naive baselines, except in the case of MAP@10 on the small dataset.

Model	**MRR**	**MAP@10**
Random Ranker	0.00094	0.00026
Popularity Based Ranker	0.03567	**0.03917**
LightGBM.LGBMRanker Baseline	0.03613	0.03774
LightGBM.LGBMRanker + Weather Features	**0.03659**	0.03866

Table 5.5: Ranking relevancy results on the small dataset

Model	**MRR**	**MAP@10**
Random Ranker	0.00044	0.00015
Popularity Based Ranker	0.04291	0.04690
LightGBM.LGBMRanker Baseline	0.04752	0.05085
LightGBM.LGBMRanker + Weather Features	**0.04816**	**0.05141**

Table 5.6: Ranking relevancy results on the large dataset

Chapter 6

Discussion

6.1 Statistical Analysis

Some of the findings from related works were also found in this statistical analysis. Several possible correlations between weather and sales on the Babyshop online storefronts were found. The correlations were sometimes in the direction that could be expected but sometimes in the opposite direction (e.g., casual dresses selling as a larger share in Stockholm when it is less sunny in Stockholm), which was surprising. The found correlations were not that strong, with magnitudes of Pearson and Spearman's correlation coefficients less than 0.4. The overall results indicated that all tested weather variables might be useful as features for the recommender system.

6.2 Recommender Systems

It should be noted that the popularity-based ranker performs not much worse than the LightGBM rankers. This is quite a strong naive baseline, but the results can also be indicative of poor model performance. The popular products in the last 30 days are, in most cases, still popular on the evaluation day. It can also be seen from the feature importance ranking (see Figure 5.2) that the sales trend features are among the most important features. This, along with the lack of better user-specific features (such as the users' favorite brands and categories or if the item is suitable for the age groups that the user has previously bought), can explain why the difference in the results between the popularity-based ranker and the LightGBM rankers was so small. The effect size of adding the weather features to the LightGBM ranker when compared to the LightGBM baseline is small. This is also in line with the findings in the statistical analysis, where the correlations between sales and weather are weak. Other factors besides the features used for the model likely affect the customers' purchasing decisions, for example, personal taste or planned use case of the product. This helps to explain why the improvement in ranking relevancy when adding the weather features was not larger.

This evaluation is performed on historical data only. To know if a recommender system with weather features would outperform the model used in a production setting, one should ideally also perform an online evaluation through A/B testing or similar.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we investigated how the performance of an ML-based recommender system differs with or without weather features in the domain of ranking children's clothing in e-commerce. First, possible correlations between local weather variables and sales on the Babyshop platforms were analyzed. All the analyzed weather variables were then turned into features that were used along with sales features and sales trends features for a LightGBM ranker. When evaluating the rankers on ranking relevancy, the LightGBMRanker + weather features outperformed the baseline with 1.11 - 2.23% across all configurations and outperformed the item popularity-based ranker in three out of four configurations. Other factors not captured by the features used in the LightGBM ranker model likely also affect the customers' purchasing decisions. This could explain why the effect size of adding the weather features is quite small.

7.2 Limitations

The original version of the recommender system program was poorly optimized, which meant that it took a long time and a lot of computing to test the program and obtain results. Eventually, a significant amount of time was spent optimizing it, which decreased the RAM load and made it run faster. Since no user features were used besides the city, many variables only needed to be calculated once for Stockholm, Göteborg, and Malmö per split rather than recomputing them for each user. By saving and fetching these variables, the program was optimized.

The lack of user-specific features likely limited the result of the LightGBM rankers. These features are expensive to compute (since they have to be calculated for every user) and would have required more time to be spent on the coding.

Since we only evaluated the recommender systems on ranking relevancy, it is possible that the baseline LightGBM ranker outperforms the ranker with weather features on other objectives (see section 2.2.3).

Regarding the statistical analysis, more effort could have been made to quantify the results and perform more types of statistical tests. It could also have been possible to combine multiple weather parameters and create multiple linear regression models besides simple linear regression models.

7.3 Future work

The weather data used in this thesis only consisted of historical observational data. Using historical weather forecasts as data could be interesting in the case that shoppers are aware of the weather forecasts and are influenced by this in their shopping decisions.

Obtaining local weather data uniquely for every customer becomes costly at scale. For this reason, a single weather data observation source was used to include customers up to about a 25 km distance from the observation point. It would be interesting to study if the ranking quality is affected by the coarseness of the weather data, and how a potential tradeoff between ranking quality and cost of weather data collection could be minimized

In the LightGBM Ranker training, other sampling methods besides random negative sampling, such as popularity-biased negative sampling [68] could have been tried to possibly improve the ranking performance. The evaluation of the recommender systems could be performed on different sales data than the Babyshop sales data, in more cities than just Stockholm, Göteborg, and Malmö and in more countries than only Sweden.

References

[1] eMarketer, "Number of digital buyers worldwide from 2014 to 2021," https://www-statista-com.focus.lib.kth.se/statistics/251666/number-of-digital-buyers-worldwide/, Jul. 2017. [Page 1.]

[2] ——, "E-commerce as percentage of total retail sales worldwide from 2015 to 2027," https://www-statista-com.focus.lib.kth.se/statistics/534123/e-commerce-share-of-retail-sales-worldwide/, Jul. 2022. [Page 1.]

[3] Dynata, "Benefits of e-commerce among global consumers as of February 2022," https://www-statista-com.focus.lib.kth.se/statistics/1308162/online-shopping-benefits-worldwide/, Mar. 2022. [Page 1.]

[4] I. Katsov, *Introduction to Algorithmic Marketing: Artficial Intelligence for Marketing Operations*. Sunnyvale, Calivornia: Ilya Katsov, 2018. ISBN 978-0-692-98904-3 [Pages 1, 12, 13, and 14.]

[5] *WMO Statement on the Status of the Global Climate in 2012*. Geneva: World Meteorological Organization, 2013. ISBN 978-92-63-11108-1 [Page 2.]

[6] V. Martínez-de-Albéniz and A. Belkaid, "Here comes the sun: Fashion goods retailing under weather fluctuations," *European Journal of Operational Research*, vol. 294, no. 3, pp. 820–830, Nov. 2021. doi: 10.1016/j.ejor.2020.01.064 [Pages 2, 18, and 39.]

[7] Y. Bahng and D. H. Kincade, "The relationship between temperature and sales: Sales data analysis of a retailer of branded women's business wear," *International Journal of Retail & Distribution Management*, vol. 40, no. 6, pp. 410–426, May 2012. doi: 10.1108/09590551211230232 [Pages 2 and 18.]

[8] “Recommender Systems in Fashion & Retail,” https://recsys.acm.org/recsys22/fashionxrecsys/. [Page 3.]

[9] A. Vishnevskiy, “GBM Ranking,” https://www.kaggle.com/code/alexvishnevskiy/gbm-ranking, 2022. [Pages 5 and 35.]

[10] C. García Ling and J. Ferrando, “H&M Personalized Fashion Recommendations,” https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations, 2022. [Page 5.]

[11] “RecSys Challenge Winners,” https://recsys.acm.org/challenges/. [Page 5.]

[12] W. Fisher, “On Grouping for Maximum Homogeneity,” *Journal of the American Statistical Association*, vol. 158, no. 53(284), pp. 789–798, 1958. doi: 10.2307/2281952 [Page 5.]

[13] “Categorical Feature Support,” https://lightgbm.readthedocs.io/en/latest/Advanced-Topics.html#categorical-feature-support. [Page 5.]

[14] Microsoft, “Optimal Split for Categorical Features,” lightgbm.readthedocs.io/en/latest/Features.html#optimal-split-for-categorical-features, 2023. [Pages 5 and 32.]

[15] M. Corporation, “Light Gradient Boosting Machine,” https://github.com/microsoft/LightGBM. [Page 5.]

[16] “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” scholar.google.se/scholar?q=LightGBM:+A+Highly+Efficient+Gradient+Boosting+Decision+Tree [Page 5.]

[17] “How to query weather forecasts from the past – Historical Forecasts,” https://www.visualcrossing.com/resources/documentation/weather-data/how-to-query-weather-forecasts-from-the-past-historical-forecasts/, Sep. 2021. [Page 6.]

[18] “Environmental Sustainability in the Fashion Industry,” https://www.genevaenvironmentnetwork.org/resources/updates/sustainable-fashion/, Jul. 2023. [Page 6.]

[19] A. Collins, D. Tkaczyk, A. Aizawa, and J. Beel, “A Study of Position Bias in Digital Library Recommender Systems,” *arXiv preprint arXiv:1802.06565*, 2018. [Page 7.]

[20] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *AI & SOCIETY*, vol. 35, no. 4, pp. 957–967, Dec. 2020. doi: 10.1007/s00146-020-00950-y [Page 7.]

[21] P. Newbold, W. L. Carlson, and B. Thorne, *Statistics for Business and Economics*, 8th ed. Boston: Pearson, 2013. ISBN 978-0-13-274565-9 [Pages 9, 10, and 11.]

[22] G. L. Shevlyakov and H. Oja, *Robust Correlation: Theory and Applications*, 3rd ed., ser. Wiley Series in Probability and Statistics. Chichester, West Sussex, United Kingdom: Wiley, 2016. ISBN 978-1-119-26453-8 978-1-119-26449-1 [Page 10.]

[23] C. C. Aggarwal, *Recommender Systems*. Cham: Springer International Publishing, 2016. ISBN 978-3-319-29657-9 978-3-319-29659-3 [Page 13.]

[24] S. Sawtelle, "Mean Average Precision (MAP) For Recommender Systems," Oct. 2016. [Page 14.]

[25] Z. Deutschman, "Recommender Systems: Machine Learning Metrics and Business Metrics - MLOps Blog," Apr. 2023. [Page 14.]

[26] S. M. Beitzel, E. C. Jensen, and O. Frieder, "MAP," *Encyclopedia of Database Systems*, 2009. [Page 14.]

[27] X. Wang, C. Li, N. Golbandi, M. Bendersky, and M. Najork, "The LambdaLoss Framework for Ranking Metric Optimization," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Torino Italy: ACM, Oct. 2018. doi: 10.1145/3269206.3271784. ISBN 978-1-4503-6014-2 pp. 1313–1322. [Pages 15 and 16.]

[28] T.-Y. Liu, *Learning to Rank for Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. ISBN 978-3-642-14266-6 978-3-642-14267-3 [Page 15.]

[29] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, Oct. 2001. doi: 10.1214/aos/1013203451 [Page 15.]

[30] A. Mohan, Z. Chen, and K. Weinberger, "Web-Search Ranking with Initialized Gradient Boosted Regression Trees," *Proceedings of the 2010*

International Conference on Yahoo! Learning to Rank Challenge - Volume 14, Jun. 2010. [Page 15.]

[31] R. Johansson, "An intuitive explanation of gradient boosting." [Page 15.]

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *NIPS'17*, Dec. 2017, pp. 3149–3157. [Pages 15 and 16.]

[33] C. J. Burges, R. Ragno, and Q. V. Le, "Learning to Rank with Nonsmooth Cost Functions," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hofmann, Eds. The MIT Press, Sep. 2007, pp. 193–200. ISBN 978-0-262-25691-9 [Page 16.]

[34] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An Overview," *Microsoft Research Technical Report MSR-TR-2010-82*, 2010. [Page 16.]

[35] Microsoft, "Lightgbm.LGBMRanker," lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRanker.html, 2023. [Pages 16 and 42.]

[36] S. Steinker, K. Hoberg, and U. W. Thonemann, "The Value of Weather Information for E-Commerce Operations," *Production and Operations Management*, vol. 26, no. 10, pp. 1854–1874, Oct. 2017. doi: 10.1111/poms.12721 [Page 17.]

[37] F. Badorf and K. Hoberg, "The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores," *Journal of Retailing and Consumer Services*, vol. 52, p. 101921, Jan. 2020. doi: 10.1016/j.jretconser.2019.101921 [Page 17.]

[38] J. Oh, Y.-H. Jo, and K.-J. Ha, "The effect of anomalous weather on the seasonal clothing market in New York," *Meteorological Applications*, vol. 28, no. 2, Mar. 2021. doi: 10.1002/met.1982 [Page 18.]

[39] J. Oh, K.-J. Ha, and Y.-H. Jo, "A Predictive Model of Seasonal Clothing Demand with Weather Factors," *Asia-Pacific Journal of Atmospheric Sciences*, vol. 58, no. 5, pp. 667–678, Dec. 2022. doi: 10.1007/s13143-022-00284-3 [Page 18.]

[40] B. S. Hoon, H. S. Chan, Jun-Ki Hong, Ji-Yeon Oh, and Ji-Su Lee, "Sales Volume Prediction Model for Temperature Change using Big Data

Analysis," *The Korea Journal of BigData*, vol. 4, no. 1, pp. 29–38, Aug. 2019. doi: 10.36498/KBIGDT.2019.4.1.29 [Pages 18, 20, and 39.]

[41] Jun-Ki Hong, "Analysis of Sales Volume by Products According to Temperature Change Using Big Data Analysis," *The Korea Journal of BigData*, vol. 4, no. 2, pp. 85–91, Dec. 2019. doi: 10.36498/KBIGDT.2019.4.2.85 [Pages 18 and 20.]

[42] E. H. Kim, H. Hwangbo, and J. M. Chae, "The effects of meteorological factors on the sales volume of apparel products - Focused on the Fall/Winter season -," *The Research Journal of the Costume Culture*, vol. 25, no. 2, pp. 117–129, Apr. 2017. doi: 10.7741/RJCC.2017.25.2.117 [Pages 19 and 20.]

[43] J. H. Oh, H. S. Oh, and K. M. Choi, "A Study on Clothes Sales Forecast System using Weather Information: Focused on S/S Clothes," *Fashion & Textile Research Journal*, vol. 19, no. 3, pp. 289–295, Jun. 2017. doi: 10.5805/SFTI.2017.19.3.289 [Pages 19 and 20.]

[44] J. Tian, Y. Zhang, and C. Zhang, "Predicting consumer variety-seeking through weather data analytics," *Electronic Commerce Research and Applications*, vol. 28, pp. 194–207, Mar. 2018. doi: 10.1016/j.elerap.2018.02.001 [Page 19.]

[45] K. B. Murray, F. Di Muro, A. Finn, and P. Popkowski Leszczyc, "The effect of weather on consumer spending," *Journal of Retailing and Consumer Services*, vol. 17, no. 6, pp. 512–520, Nov. 2010. doi: 10.1016/j.jretconser.2010.08.006 [Page 20.]

[46] J. Oh, K.-J. Ha, and Y.-H. Jo, "Use of Weather Factors in Clothing Studies in Korea and its Implications: A Review," *Asia-Pacific Journal of Atmospheric Sciences*, vol. 58, no. 5, pp. 729–741, Dec. 2022. doi: 10.1007/s13143-022-00279-0 [Page 20.]

[47] J. Hong, H. Lee, and J.-H. Na, "Effects of meteorological factors on the sales of seasonal products," Tech. Rep. 15-6, 2012. [Pages 20 and 39.]

[48] YK. Lee, KH. Ahn, and N. Chung, "An effect of weather on firm's sales on clothes shop," Tech. Rep. 13-1, 2011. [Page 20.]

[49] E.-Y. Jang and B.-H. Lim, "An exploratory study on the effect of weather factors on sales of fashion apparel products in department stores," Tech. Rep. 12, 2003. [Page 20.]

[50] Y. Liu, Y. Gao, S. Feng, and Z. Li, "Weather-to-garment: Weather-oriented clothing recommendation," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. Hong Kong, Hong Kong: IEEE, Jul. 2017. doi: 10.1109/ICME.2017.8019476. ISBN 978-1-5090-6067-2 pp. 181–186. [Page 20.]

[51] Y. Wen, X. Liu, and B. Xu, "Personalized Clothing Recommendation Based on Knowledge Graph," in *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. Shanghai: IEEE, Jul. 2018. doi: 10.1109/ICALIP.2018.8455311. ISBN 978-1-5386-5195-7 pp. 1–5. [Page 21.]

[52] L. Yu-Chu, Y. Kawakita, E. Suzuki, and H. Ichikawa, "Personalized Clothing-Recommendation System Based on a Modified Bayesian Network," in *2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet*. Izmir, Turkey: IEEE, Jul. 2012. doi: 10.1109/SAINT.2012.75. ISBN 978-1-4673-2001-6 978-0-7695-4737-4 pp. 414–417. [Page 21.]

[53] Hao Peifeng, Cui Yuzhe, Song Jingping, and Hu Zhaomu, "Smart wardrobe system based on Android platform," in *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. Chengdu, China: IEEE, Jul. 2016. doi: 10.1109/ICCCBDA.2016.7529571. ISBN 978-1-5090-2594-7 pp. 279–285. [Page 21.]

[54] R. Koshy, A. Gharat, T. Wagh, and S. Sonawane, "A Complexion based Outfit color recommender using Neural Networks," in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. Bhilai, India: IEEE, Feb. 2021. doi: 10.1109/ICAECT49130.2021.9392418. ISBN 978-1-72815-791-7 pp. 1–7. [Page 21.]

[55] M. C. Keller, B. L. Fredrickson, O. Ybarra, S. Côté, K. Johnson, J. Mikels, A. Conway, and T. Wager, "A Warm Heart and a Clear Head: The Contingent Effects of Weather on Mood and Cognition," *Psychological Science*, vol. 16, no. 9, pp. 724–731, Sep. 2005. doi: 10.1111/j.1467-9280.2005.01602.x [Page 21.]

[56] Y. Zwebner, L. Lee, and J. Goldenberg, "The temperature premium: Warm temperatures increase product valuation," *Journal of*

Consumer Psychology, vol. 24, no. 2, pp. 251–259, Apr. 2014. doi: 10.1016/j.jcps.2013.11.003 [Page 21.]

[57] A. Bassi, R. Colacito, and P. Fulghieri, "'O Sole Mio: An Experimental Analysis of Weather and Risk Attitudes in Financial Decisions," *Review of Financial Studies*, vol. 26, no. 7, pp. 1824–1852, Jul. 2013. doi: 10.1093/rfs/hht004 [Page 21.]

[58] L. Buchheim and T. Kolaska, "Weather and the Psychology of Purchasing Outdoor Movie Tickets," *Management Science*, vol. 63, no. 11, pp. 3718–3738, Nov. 2017. doi: 10.1287/mnsc.2016.2524 [Page 22.]

[59] "Datetimes and Timedeltas," numpy.org/doc/stable/reference/arrays.datetime. [Page 25.]

[60] "Measure Circle / Radius on a map," https://www.maps.ie/draw-radius-circle-map/, 2023. [Page 25.]

[61] "Postorter i kommun-ordning," http://www.post24.se/postort-kommun-lan-kommunordning/. [Page 25.]

[62] "Easy Global Weather API - Single history & forecast Weather API," https://www.visualcrossing.com/weather-api. [Page 26.]

[63] "Modifying the weather station search parameters," https://www.visualcrossing.com/resources/documentation/weather-data/modifying-the-weather-station-search-parameters/, Apr. 2023. [Pages 26 and 27.]

[64] "Weather Data Documentation," visualcrossing.com/resources/documentation/weather-data/weather-data-documentation. [Page 27.]

[65] I. Lorentz, "KTH DA233X Machine Learning Master Thesis Repo," github.com/IsacLorentz/MSc-thesis-LightGBMRanker-weather-features-fashion, Aug. 2023. [Page 27.]

[66] S. Iyyer, "Step by Step Assumptions - Linear Regression," https://www.kaggle.com/code/shrutimechlearn/step-by-step-assumptions-linear-regression, Jun. 2020. [Page 32.]

[67] B. Hamner, "Metrics/Python/ml_metrics /average_precision.py," github.com/benhamner/Metrics/blob/master/Python/ml_metrics/average_precision.py, Aug. 2015. [Page 37.]

[68] B. Liu and B. Wang, “Bayesian Negative Sampling for Recommendation,” Jul. 2022, comment: 21 pages. [Page 48.]

Appendix A

Sales Data Muncipalities and Postal Towns

City	Muncipalities
Stockholm	Botkyrka, Danderyd, Ekerö, Haninge, Huddinge, Järfälla, Lidingö, Nacka, Nykvarn, Nynäshamn, Salem, Sigtuna, Sollentuna, Solna, Sthlm, Stockholm, Sundyberg, Södertälje, Tyresö, Täby, Upplands Väsby, Upplands-Bro, Vallentuna, Vaxholm, Värmdö, Österåker
Göteborg	Ale, Alingsås, Bollebygd, Göteborg, Härryda, Kungsbacka, Kungälv, Lerum, Lerum, Lilla edet, Mölndal, Partille, Stenungsund, Stenungsund, Tjörn, Öckerö
Malmö	Burlöv, Eslöv, Kävlinge, Landskrona, Lomma, Lund, Malmö, Skurup, Staffanstorp, Svalöv, Svedala, Trelleborg, Vellinge

Table A.1: Sales data municipalities

List A.1: Sales data postal cities

- **Stockholm**: Adelsö, Arlandastad, Bagarmossen, Bandhagen, Brandbergen, Bro, Bromma, Brottby, Dalarö, Danderyd, Djurhamn, Djursholm, Drottningholm, Ekerö, Enebyberg, Enhörna, Enskede, Enskede Gård, Enskededalen, Farsta, Färentuna, Grinda, Gränö, Grödinge, Gustavsberg, Gålö, Gällnöby, Handen, Haninge, Harö, Huddinge, Husarö, Hårsfjärden, Hägersten, Hässelby, Hölö, Ingarö, Ingmarsö, Järfälla, Järna, Johanneshov, Jordbro, Kista, Kungens Kurva, Kungsängen, Lidingö, Ljusterö, Märsta, Märsta Arlanda, Möja, Mölnbo, Mörkö, Munsö, Muskö, Nämdö, Nacka, Nacka strand, Norra Sorunda, Norrby, Norsborg, Nykvarn, Nynäshamn, Ornö, Rosersberg, Runmarö, Rönninge, Saltsjö-Boo, Saltsjö-Duvnäs, Saltsjöbaden, Sandhamn, Segeltorp, Segersäng, Sigtuna, Skå, Skälvik, Skärholmen, Sköndal, Skarpnäck, Skogås, Sollenkroka Ö, Sollentuna, Solna, Sorunda, Spånga, Stavsnäs, Stavsudda, Stenhamra, Steningehöjden, Stockholm, Stockholm-Arlanda, Stockholm-Globen, Stocksund, Stora Vika, Svartsjö, Söderby, Södertälje, Tomteboda, Trångsund, Tullinge, Tumba, Tungelsta, Tyresö, Täby, Upplands Väsby, Uttran, Utö, Vallentuna, Vaxholm, Vega, Vendelsö, Vårby, Vällingby, Värmdö, Västerhaninge, Åkersberga, Årsta, Årsta havsbad, Älta, Älvsjö, Ösmo, Österhaninge, Österskär

- **Göteborg:** Agnesberg, Alafors, Alingsås, Alvhem, Angered, Askim, Asperö, Billdal, Bohus, Bohus-Björkö, Bollebygd, Brännö, Donsö, Fjärås, Floda, Fotö, Frillesås, Göta, Göteborg, Gråbo, Grötö, Gunnilse, Harestad, Hindås, Hisings Backa, Hisings kärra, Hjärtum, Hovås, Hultafors, Hyppeln, Hålanda, Hålta, Hällingsjö, Hälsö, Härryda, Hönö, Jörlanda, Jonsered, Kållered, Källö-Knippla, Kärna, Köpstadsö, Kalvsund, Kareby, Kode, Kullavik, Kungälv, Kungsbacka, Lödöse, Landvetter, Lerum, Lilla Edet, Lindome, Lycke, Mölndal, Mölnlycke, Marstrand, Nödinge, Nol, Nygård, Olofstorp, Olsfors, Onsala, Partille, Pixbo, Prässebo, Romelanda, Rävlanda, Rörö, Sjövik, Skepplanda, Sollebrunn, Spekeröd, Stenkullen, Stenungsund, Stora Höga, Styrsö, Surte, Svenshögen, Särö, Säve, Sävedalen, Tollered, Torslanda, Töllsjö, Ucklum, Vallda, Vrångö, Västerlanda, Västra frölunda, Ytterby, Åsa, Älvängen, Öckerö, Ödsmål, Öjersjö

- **Malmö:** Abbekås, Alnarp, Anderslöv, Annelöv, Arlöv, Asmundtorp, Bara, Barsebäck, Beddingestrand, Billeberga, Billinge, Bjärred, Bunkeflostrand, Dösjebro, Dalby, Eslöv, Falsterbo, Flyinge, Furulund, Genarp, Glumslöv, Harlösa, Häljarp, Härslöv, Höllviken, Kågeröd, Kävlinge, Klågerup, Klagshamn, Klagstorp, Löberöd, Löddeköpinge, Landskrona, Limhamn, Lomma, Lund, Malmö, Malmösturup, Marieholm, Oxie, Rydsgård, Röstånga, Sankt ibb, Saxtorp, Skanör, Skivarp, Skurup, Smygehamn, Stehag, Stockamöllan, Svalöv, Svedala, Södra Sandby, Teckomatorp, Torna-hällestad, Trelleborg, Tygelsjö, Tågarp, Veberöd, Vellinge, Vintrie, Åkarp

Appendix B

Used Coordinates for City Centers

City	Latitude	Longitude
Göteborg	57.7072399861	11.9670055611
Malmö	55.6053338582	13.0001149927
Stockholm	59.3251232467	18.0710393336

Table B.1: Used coordinates for centers of Göteborg, Malmö and Stockholm

Appendix C

Simple Moving Averages Pandas Code

The simple moving averages were computed this way where *stockholm_sales_delta* is a pandas.DataFrame:

```
stockholm_sales_delta[f"{category}_share_difference_percent_SMA_{sma}"] = (
        stockholm_sales_delta[f"{category}_share_difference_percent"]
        .rolling(window=f"{sma}D", min_periods=sma)
        .mean()
    )

stockholm_sales_delta[f"{weather_param}_difference_SMA_{sma}"] = (
    stockholm_sales_delta[f"{weather_param}_difference"]
    .rolling(window=f"{sma}D", min_periods=sma)
    .mean()
)

stockholm_sales_delta = stockholm_sales_delta.dropna(
    subset=[
        f"{category}_share_difference_percent_SMA_{sma}",
        f"{weather_param}_difference_SMA_{sma}",
    ]
)
```

Listing C.1: Simple Moving Average Data Processing

Appendix D

Statistical Analysis Supporting Results

D.1 Results From a Stable Positive Correlation: Strappy Sandals and Feelslike Temperature

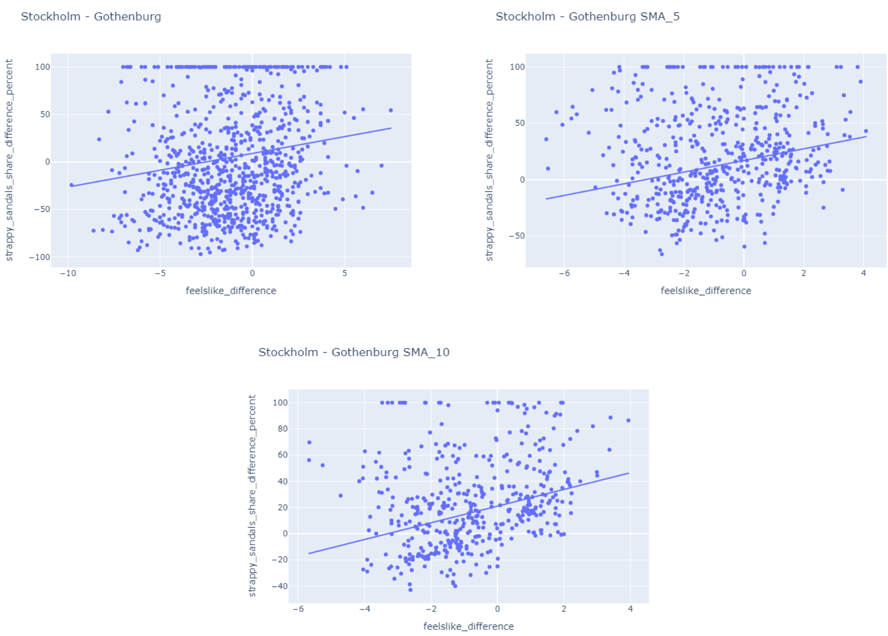

Figure D.1: Figures showing the linear regression for daily values, 5 days simple moving average and 10 days simple moving average

Configuration	Pearson correlation coefficient (p-value)	Spearman's rank correlation coefficient (p-value)
Daily values	0.2091 (2.43e-259)	0.2392 (0.0)
5 days moving average	0.2706 (0.0)	0.3440 (0.0)
10 days moving average	**0.3155 (0.0)**	**0.3918 (0.0)**

Table D.1: Pearson and Spearman's correlation coefficients per configuration

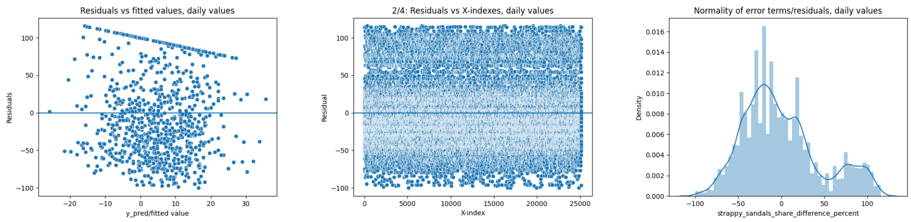

Figure D.2: Test for assumptions of linear regression, daily values

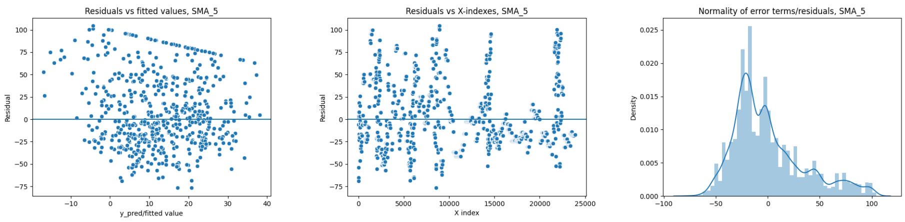

Figure D.3: Test for assumptions of linear regression, 5 days simple moving average values

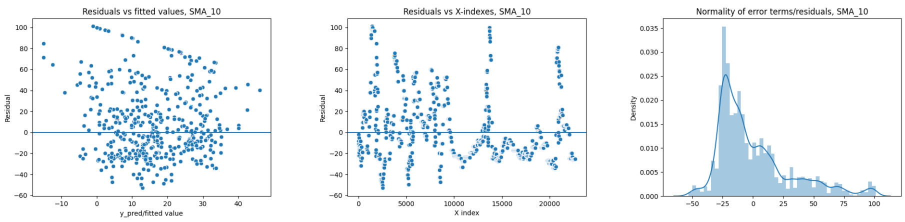

Figure D.4: Test for assumptions of linear regression, 10 days simple moving average values

Configuration	Mean of residuals
Daily values	-2.5327-16
5 days moving average	3.0405e-15
10 days moving average	-9.6276e-16

Table D.2: Mean of residuals per configuration (should be 0)

D.2 Results From No Correlation Found: Casual Dresses and Feels-Like Temperature

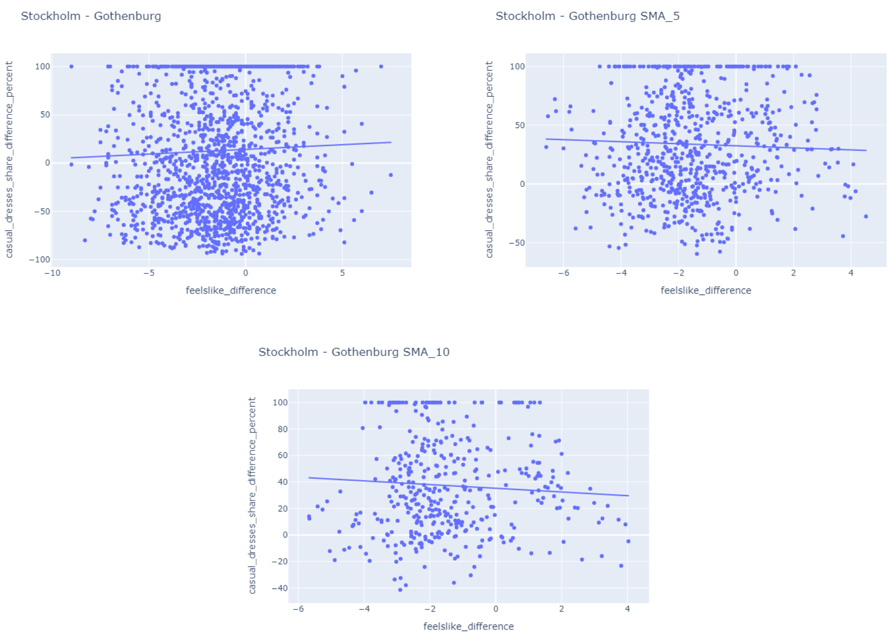

Figure D.5: Figures showing the linear regression for daily values, 5 days simple moving average and 10 days simple moving average

Configuration	Pearson correlation coefficient (p-value)	Spearman's rank correlation coefficient (p-value)
Daily values	0.0428 (1.2243e-11)	0.0458 (4.0577e-13)
5 days moving average	-0.0386 (1.8877e-6)	-0.0180 (0.0256)
10 days moving average	-0.0622 (2.9778e-9)	-0.0152 (0.1491)

Table D.3: Pearson and Spearman's correlation coefficients per configuration

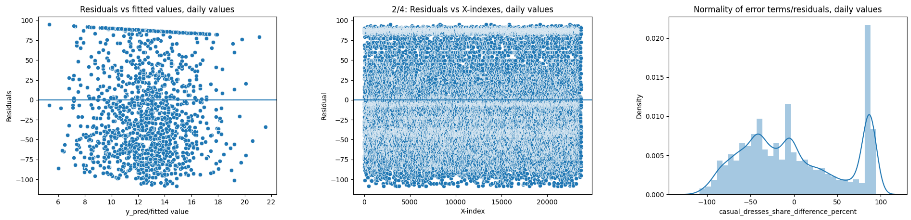

Figure D.6: Test for assumptions of linear regression, daily values

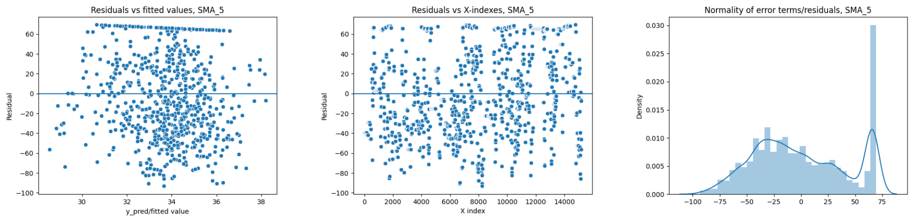

Figure D.7: Test for assumptions of linear regression, 5 days simple moving average values

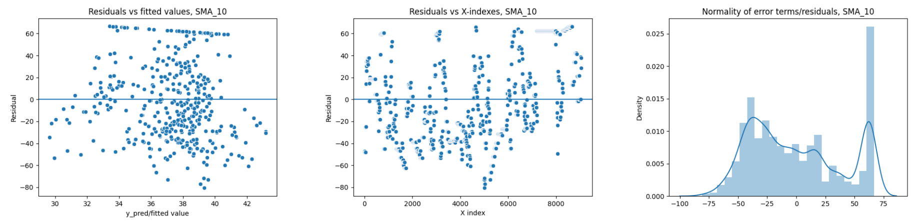

Figure D.8: Test for assumptions of linear regression, 10 days simple moving average values

Configuration	**Mean of residuals**
Daily values	-1.5346e-16
5 days moving average	-9.5780e-16
10 days moving average	5.6203e-15

Table D.4: Mean of residuals per configuration (should be 0)

D.3 Results From a Stable Negative Correlation: Rain Sets and Solar Energy

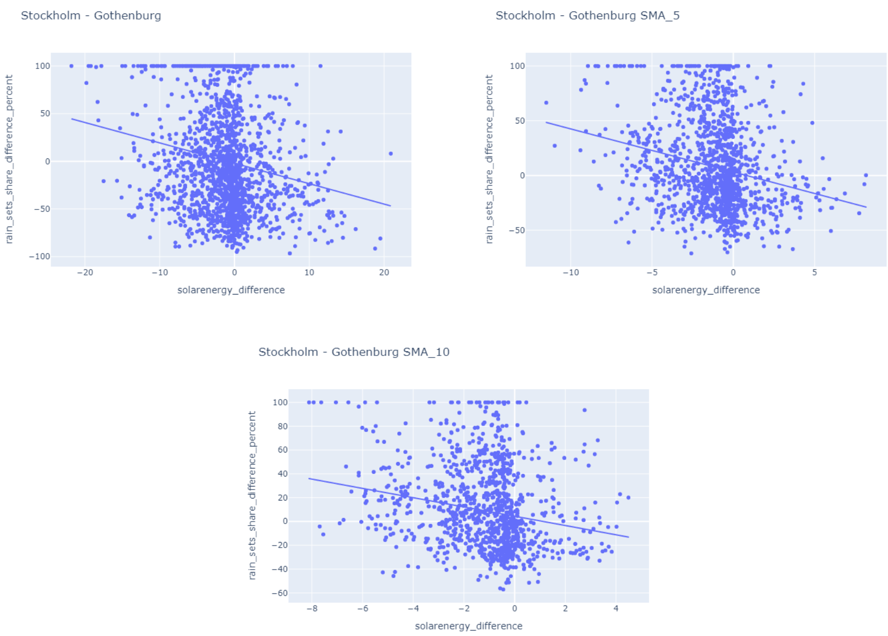

Figure D.9: Figures showing the linear regression for daily values, 5 days simple moving average and 10 days simple moving average

Configuration	Pearson correlation coefficient (p-value)	Spearman's rank correlation coefficient (p-value)
Daily values	-0.2084 (0.0)	-0.2007 (0.0)
5 days moving average	**-0.2307 (0.0)**	-0.2444 (0.0)
10 days moving average	-0.2078 (0.0)	**-0.2593 (0.0)**

Table D.5: Pearson and Spearmanäs correlation coefficients per configuration

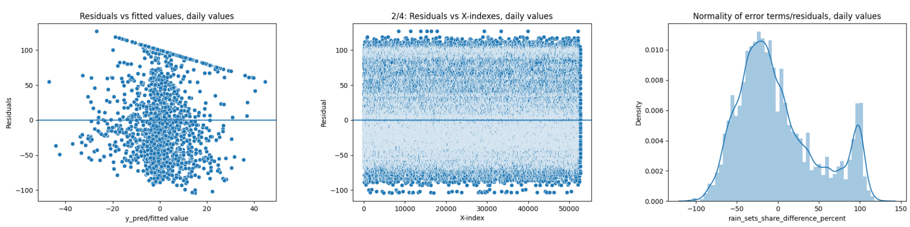

Figure D.10: Test for assumptions of linear regression, daily values

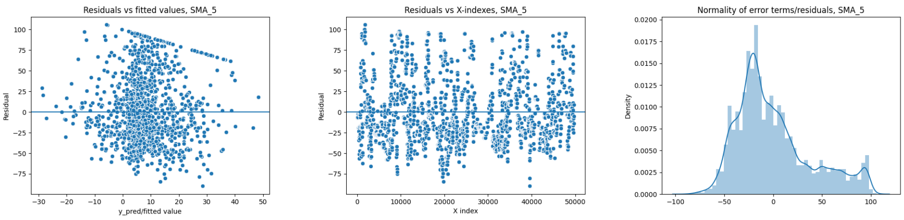

Figure D.11: Test for assumptions of linear regression, 5 days simple moving average values

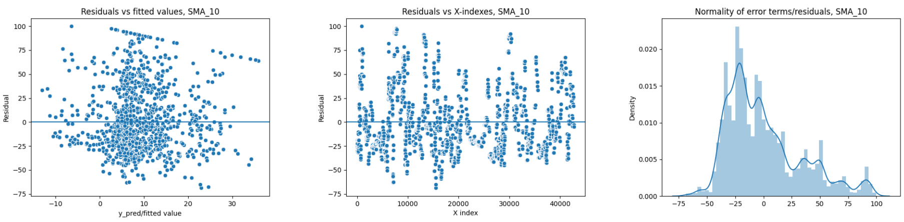

Figure D.12: Test for assumptions of linear regression, 10 days simple moving average values

Configuration	**Mean of residuals**
Daily values	-6.8995e-17
5 days moving average	-2.3023e-15
10 days moving average	-1.7845e-15

Table D.6: Mean of residuals per configuration (should be 0)

TRITA-EECS-EX- 2023:0000

€€€€ For DIVA €€€€

{
”Author1”: { ”Last name”: ”Lorentz”,
”First name”: ”Isac”,
”Local User Id”: ”ilorentz”,
”E-mail”: ”ilorentz@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
}
},
”Cycle”: ”2”,
”Course code”: ”DA233X”,
”Credits”: ”30.0”,
”Degree1”: {”Educational program”: ””
,”Degree”: ”Both Degree of Master of Science in Engineering and Master’s degree”
,”subjectArea”: ”Computer Science and Engineering, specializing in Machine Learning”
},
”Title”: {
”Main title”: ”Shoppin’ in the Rain”,
”Subtitle”: ”An Evaluation of the Usefulness of Weather-Based Features for an ML Ranking Model in the Setting of Children’s Clothing Online Retailing”,
”Language”: ”eng” },
”Alternative title”: {
”Main title”: ”Handla i regnet”,
”Subtitle”: ”En utvärdering av användbarheten av väderbaserade variabler för en ML-rankningsmodell inom onlineförsäljning av barnkläder”,
”Language”: ”swe”
},
”Supervisor1”: { ”Last name”: ”Al-Shishtawy”,
”First name”: ”Ahmad”,
”Local User Id”: ”ahmadas”,
”E-mail”: ”ahmadas@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Supervisor2”: { ”Last name”: ”Svensson”,
”First name”: ”Marcus”,
”E-mail”: ”marcus.svensson@babyshop.se”,
”Other organisation”: ”Babyshop Sthlm”
},
”Examiner1”: { ”Last name”: ”Kragic Jensfelt”,
”First name”: ”Danica”,
”Local User Id”: ”danik”,
”E-mail”: ”danik@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Intelligent Systems” }
},
”Cooperation”: { ”Partner_name”: ”Babyshop Sthlm AB”},
”National Subject Categories”: ”10201, 10207”,
”Other information”: {”Year”: ”2023”, ”Number of pages”: ”xvii,69”},
”Copyrightleft”: ”copyright”,
”Series”: { ”Title of series”: ”TRITA-EECS-EX” , ”No. in series”: ”2023:0000” },
”Opponents”: { ”Name”: ”David Yu”},
”Presentation”: { ”Date”: ”2023-09-08 13:00”
,”Language”:”eng”
,”Room”: ”via Zoom https://kth-se.zoom.us/j/62917323456”
,”Address”: ”Zoom”
,”City”: ”Stockholm” },
”Number of lang instances”: ”2”,
”Abstract[eng]”: €€€€

Online shopping offers numerous benefits, but large product catalogs make it difficult for shoppers to understand the existence and characteristics of every item for sale. To simplify the decision-making process, online retailers use ranking models to recommend products relevant to each individual user. Contextual user data, such as location, time, or local weather conditions, can serve as valuable features for ranking models, enabling personalized real-time recommendations. Little research has been published on the usefulness of weather-based features for ranking models in online clothing retailing, which makes additional research into this topic worthwhile. Using Swedish sales and customer data from Babyshop, an online retailer of children's fashion, this study examined possible correlations between local weather data and sales. This was done by comparing differences in daily weather and differences in daily shares of sold items per clothing category for two cities: Stockholm and Göteborg. With Malmö as an additional city, historical observational weather data from one location each in the three cities Stockholm, Göteborg, and Malmö was then featurized and used along with the customers' postal towns, sales features, and sales trend features to train and evaluate the ranking relevancy of a gradient boosted decision trees learning to rank LightGBM ranking model with weather features. The ranking relevancy was compared against a LightGBM baseline that omitted the weather features and a naive baseline: a popularity-based ranker. Several possible correlations between a clothing category such as shorts, rainwear, shell jackets, winter wear, and a weather variable such as feels-like temperature, solar energy, wind speed, precipitation, snow, and snow depth were found. Evaluation of the ranking relevancy was done using the mean reciprocal rank and the mean average precision @ 10 on a small dataset consisting only of customer data from the postal towns Stockholm, Göteborg, and Malmö and also on a larger dataset where customers in postal towns from larger geographical areas had their home locations approximated as Stockholm, Göteborg or Malmö. The LightGBM rankers beat the naive baseline in three out of four configurations, and the ranker with weather features outperformed the LightGBM baseline by 1.1 to 2.2 percent across all configurations. The findings can potentially help online clothing retailers create more relevant product recommendations.

€€€€,
"Keywords[eng]": €€€€
Statistical analysis, regression analysis, recommender systems, ensemble learning, electronic commerce, LightGBM, learning to rank, feature selection, weather-based features, fashion €€€€,
"Abstract[swe]": €€€€

Internethandel erbjuder flera fördelar, men stora produktsortiment gör det svårt för konsumenter att känna till existensen av och egenskaperna hos alla produkter som saluförs. För att förenkla beslutsprocessen så använder internethandlare rankningsmodeller för att rekommendera relevanta produkter till varje enskild användare. Kontextuell användardata såsom tid på dygnet, användarens plats eller lokalt väder kan vara värdefulla variabler för rankningsmodeller då det möjliggör personaliserade realtidsrekommendationer. Det finns inte mycket publicerad forskning inom nyttan av väderbaserade variabler för produktrekommendationssystem inom internethandel av kläder, vilket gör ytterligare studier inom detta område intressant. Med hjälp av svensk försäljnings- och kunddata från Babyshop, en internethandel för barnkläder så undersökte denna studie möjliga korrelationer mellan lokal väderdata och försäljning. Detta gjordes genom att jämföra skillnaderna i dagligt väder och skillnaderna i dagliga andelar av sålda artiklar per klädeskategori för två städer: Stockholm och Göteborg. Med Malmö som ytterligare en stad så gjordes historiska metereologiska observationer från en plats var i Stockholm, Göteborg och Malmö till variabler och användes tillsammans med kundernas postorter, försäljningsvariabler och variabler för försäljningstrender för att träna och utvärdera rankningsrelevansen hos en gradient-boosted decision trees learning to rank LightGBM rankningsmodell med vädervariabler. Rankningsrelevansen jämfördes mot en LightGBM baslinjesmodel som saknade vädervariabler samt en naiv baslinje: en popularitetsbaserad rankningsmodell. Flera möjliga korrelationer mellan en klädeskategori som shorts, regnkläder, skaljackor, vinterkläder och och en daglig vädervariabel som känns-som-temperatur, solenergi, vindhastighet, nederbörd, snö och snödjup upptäcktes. Utvärderingen av rankingsrelevansen utfördes med mean reciprocal rank och mean average precision @ 10 på ett mindre dataset som bestod endast av kunddata från postorterna Stockholm, Göteborg och Malmö och även på ett större dataset där kunder med postorter från större geografiska områden fick sina hemorter approximerade som Stockholm, Göteborg eller Malmö. LigthGBM-rankningsmodellerna slog den naiva baslinjen i tre av fyra konfigurationer och rankningsmodellen med vädervariabler slog LightGBM baslinjen med 1.1 till 2.2 procent i alla konfigurationer. Resultaten kan potentiellt hjälpa internethandlare inom mode att skapa bättre produktrekommendationssystem.

€€€€,
"Keywords[swe]": €€€€
Statistisk analys, regressionsanalys, rekommendationssystem, ensemble-inlärning, näthandel, LightGBM, learning to rank, variabelselektion, väderbaserade variabler, mode €€€€,
}

acronyms.tex

```
\setabbreviationstyle[acronym]{long-short}
\newacronym{GBDT}{GBDT}{Gradient Boosting Decision Trees}
\newacronym{LETOR}{LETOR}{Learning to Rank}
\newacronym{ML}{ML}{Machine Learning}
\newacronym{MAP}{MAP}{Mean Average Precision}
\newacronym{MAP@K}{MAP@K}{Mean Average Precision at K}
\newacronym{MRR}{MRR}{Mean Reciprocal Rank}
\newacronym{SMA}{SMA}{Simple Moving Average}
```