NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

DATA-DRIVEN RECOMMENDER SYSTEM FOR CAR BODYSHOP STEERING

A case study of an insurance company in Portugal

Mohamed Ettaher Ben Slama

Project Work

presented as a partial requirement for obtaining a master's degree in data science and advanced Analytics.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

DATA-DRIVEN RECOMMENDER SYSTEM FOR CAR BODYSHOP STEERING

A case study of an insurance company in Portugal

by

Mohamed Ettaher Ben Slama

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervisor: Bruno Jardim, PhD, NOVA Information Management School

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, January 2024

ACKNOWLEDGEMENT

I would like to begin by expressing my heartfelt gratitude to the entire Nova IMS community. The professors and staff, with their exceptional expertise and unwavering commitment to education, have profoundly shaped my academic journey.

I am particularly grateful to Professor Bruno Jardim, my academic supervisor. His invaluable guidance and support during my thesis work have greatly influenced the direction and quality of this research.

I also extend my deepest thanks to the Advanced Analytics and Transformation Team at Generali Tranquilidade, where I had the incredible opportunity to intern for nine months. Their mentorship has provided me with a deeper understanding of the practical applications of data science, which has been essential for my professional growth.

To my friends, thank you for your constant encouragement and support. Your belief in me has been a continuous source of motivation.

Above all, I want to express my deepest gratitude to my family. Your endless love, patience, and encouragement have been my rock. Your unwavering faith in me has been my greatest source of inspiration and strength.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CRISP-DM	Cross-Industry Standard Process for Data Mining
VIN	Vehicle Identification Number
ML	Machine Learning
DBSCAN	Density-based Spatial Clustering
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
EDA	Exploratory Data Analysis

ABSTRACT

To optimize the steering process of auto bodyshops for a Portuguese insurance company, this research proposes a data-driven recommender system. Making use of an extensive dataset that included customer evaluations, bodyshop performance measures, and claims, we used machine learning methods to create a strong ranking model to recommend most convenient bodyshop followed by clustering model to optimize the network. The work outlines the procedures from data collection to model assessment using the CRISP-DM framework. Key results show that bodyshops may be classified into separate clusters based on cost-effectiveness, customer satisfaction, and operational efficiency. The suggested approach not only improves decision-making when choosing a repair facility, but it also results in significant cost savings and higher customer satisfaction. These results demonstrate the potential of data-driven techniques in operational optimization and have wider implications for the insurance sector.

Key Words

Data-Driven Recommender System; Insurance Company; Car Bodyshops; Machine Learning; Operational Efficiency

Sustainable Development Goals (SDG):

1. INTRODUCTION

1.1 Motivation

Efficient claims management and network optimization are crucial for insurance organizations, especially considering the dynamic character of the industry (Brüggemann et al.,2018). The primary objective is to align with consumer expectations while simultaneously reducing costs to accomplish the triple objectives of customer satisfaction, cost reduction, and revenue growth. According to McKinsey & Company (2018), automation has the potential to decrease expenditures related to the claims process by up to 30%. The increasing use of AI tools such as recommender systems and predictive models (Singh & Singh, 2024) serves as evidence of the automation trend, which enhances customer loyalty and instills trust in insurers which is very important for a sustainable relationship.

Prominent insurance companies in Portugal, must carefully manage costs, increase income, and ensure client happiness, particularly when it comes to claims fees and selecting a repair facility. This project intends to improve claims costs by using a data-driven recommender system to attain operational excellence. This research was motivated by the recognition that the insurance industry has not fully adopted data-driven decision-making. Repair facility selection is a challenging issue that standard claims management approaches typically fail to handle successfully, resulting in a considerable increase in claims expenses. Data analytics can revolutionize the insurance industry by improving operational efficiency, providing customers with more precise recommendations, and driving continuous advancements in the field.

1.2 Steering process optimization

Typically, once an accident occurs, the process of guiding consumers to a garage/bodyshop involves calling the insurance company. When a consumer seeks recommendations for repair garages, they often contact their insurance carrier to file a claim. Optimizing consumer guidance techniques to direct consumers to certain bodyshops is a challenging issue that involves considering factors such as repair pricing, customer preferences, and geographic locations. Nevertheless, these factors are considered in the current process. It continues to depend on human input and lacks the educational aspect of historical data analytics. Utilizing data science and advanced analytics may significantly enhance the effectiveness and efficiency of this operation, aligning it with the primary goals of enhancing customer satisfaction and minimizing claims costs.

The study conducted by Smith and Johnson (Smith & Johnson, 2020) has shown the potential advantages of using sophisticated analytics and data science to optimize customer support processes. These methods have proven effective in enhancing customer satisfaction and optimizing operational productivity. Another study highlighted the potential of advanced

analytics in minimizing claims costs by improving decision-making in selecting bodyshops (Lee et al., 2019). These results emphasize the need for using data-driven approaches to enhance customer satisfaction and reduce the cost of referring consumers to appropriate bodyshops. Implementing these innovative solutions may enhance decision-making, streamline the steering process, and ultimately achieve the objectives of enhancing customer satisfaction and minimizing costs when sending clients to appropriate bodyshops.

1.3 The case of Portuguese Insurance

Typically, garage recommendations in Portugal are sent by telephone with an operator. The operator employs a sophisticated decision-making process by carefully considering several factors, such as brand specialization, network affiliation, and geographical location. Despite its beauty, the human touch typically encounters a barrier in terms of a relatively low success rate. A good method for demonstrating the advantages and challenges of claims cost optimization is by using a typical Portuguese insurance company as an example. This case study examines the current processes used by an insurance company and highlights ongoing issues. Furthermore, it provides a valuable framework for understanding the difficulties of implementing a data-driven recommendation system in a practical insurance environment.

1.4 Objectives and methodology

This study aims to create a recommender system that enhances the selection of preferred bodyshops for insurance clients. It also proposes practical adjustments to terms and conditions by analyzing data and considering consumer preferences. It utilizes the inspiration and insights acquired from an in-depth analysis of a representative Portuguese insurance firm. Methodologically, the study adopts a comprehensive approach:

1. Data Collection: Diverse data will be sourced, including claims data, garage performance metrics, expert evaluations, and external information about garages and vehicles.
2. Predictive Models: Utilizing statistical analysis and predictive modeling, iterative development will lead to the mathematical optimization of repair shop costs.

This study aims to explore the theoretical underpinnings, address methodological challenges, and provide valuable insights into the development and operation of a data-driven recommendation system tailored for a Portuguese insurance company, following the guidelines of the Cross-Industry Standard Process for Data Mining (CRISP-DM). The project consists of the following components:

- Literature Review: This section delves into core concepts and combines pertinent research, academic papers, and theoretical frameworks to facilitate a research study.
- Methodology: Describes the process of collecting data, clustering for car bodyshops, the models used, and the interpretation

• Results and Discussion: This section assesses and deliberates on the findings and their broader implications, and practical business applications.
• Conclusions: Revive the work and important findings, emphasizing the importance of the inquiry while acknowledging the study's limitations and the project's objectives.
• Limitations and Recommendations for Future Projects: highlights the difficulties faced throughout the research and provides recommendations for other areas of investigation

2. LITERATURE REVIEW

This section provides a comprehensive theoretical foundation for the project, beginning with a thorough overview of Recommender Systems (Section 2.1), detailing their principles, types, and applications. An overview of Machine Learning follows (Section 2.2), explaining fundamental concepts, types of learning, and key algorithms. Sentiment analysis is then discussed (Section 2.3), covering techniques and applications for understanding user opinions. Finally, Web Scraping is explored (Section 2.4), discussing its relevance in data collection.

2.1 Recommender systems

The Internet became widely accessible for data retrieval purposes in the early 1990s. The explosion escalated into a predicament that needed immediate action. Many websites that provide a wide range of information, including news items, articles, and products for sale, have been shown to have problems with their customer support. Recommender Systems, as proposed by Resnick and Varian in 1997, provide a curated selection of goods that are likely to be of interest to a certain user, hence facilitating the user's process of product discovery. Recommender systems, also known as recommendation systems, are information filtering systems that use analytics and data mining to extract pertinent information from vast amounts of available data. These systems analyze user behaviors, activities, and preferences to provide appropriate recommendations. Recommendation systems, which suggest various forms of content such as books, music, movies, videos, and newspapers, have gained significant significance in the era of big data and are now often used in daily life. (Zhang et al., 2023) Recommender systems were first developed using conventional data mining techniques such as association rules. Subsequent undertakings, despite initial setbacks, have yielded significant outcomes that may be classified into three distinct groups. Collaborative remodels and content-based recommender systems (Shani et al., 2005). Recommender systems are often categorized into two groups: computer systems and hybrid recommender systems. Figure 1 presents a schematic representation of the many classifications of recommender systems.

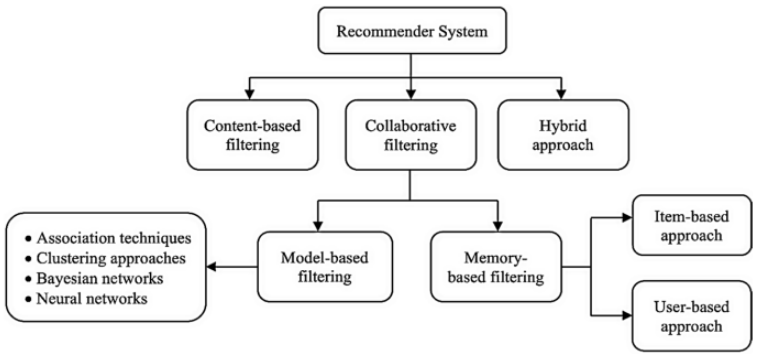

Figure 1 - Types of recommender systems. Taken from (Dutta & Roy, 2022).

2.2.1 Content-based recommender system.

Data items in content-based recommender systems are organized into item profiles depending on their features or descriptions. Some features of a book may include the author, publisher, and other relevant details. When discussing a movie, its features include the director, actor, and other relevant individuals (Dutta & Roy, 2022). The rating of an item is merged with the favorable ratings of its other components to generate a user profile. This user profile encompasses all items.

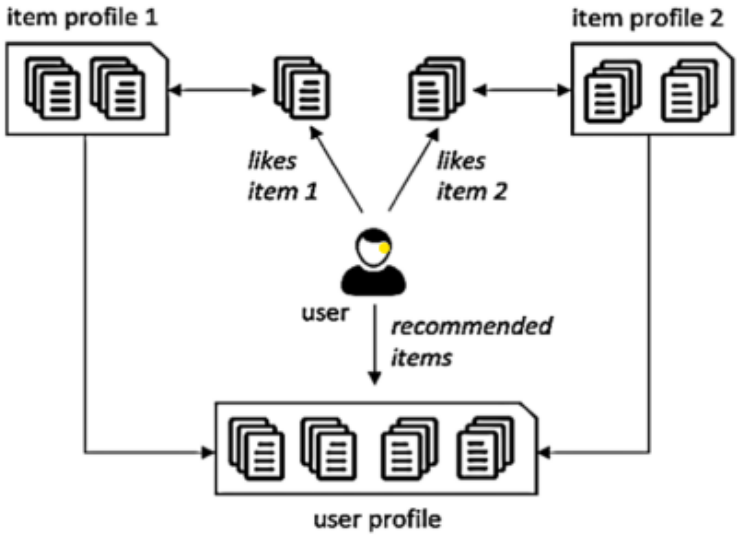

Figure 2- Content-based recommender system. Taken from (Dutta & Roy, 2022).

Profiles, whose items are rated positively by the user. Items present in this user profile are then recommended to the user, as shown in Figure 2. Content-based filtering uses item features to recommend other items like what the user likes, based on their previous actions or explicit feedback (Roy & Dutta, 2022). The model should recommend items relevant to this user. To do so, you must first pick a similarity metric (for example, dot product). Then, you must set up the system to score each candidate item according to this similarity metric. Note that the recommendations are specific to this user, as the model did not use any information about other users (Roy & Dutta, 2022). Using Dot Product as a Similarity Measure: Consider the case of a book where the reader embedding X and the book embedding Y are both binary vectors.

Since $<X,Y>=i=0n\ X_i\ Yi$ a feature appearing in both X and Y contributes a 1 to the sum. In other words, $<X,Y>$is the number of features that are active in both vectors simultaneously.

A high dot product then indicates more common features, thus a higher similarity. Using cosine as a similarity measure: Consider the case of a book where the reader embedding X and the book embedding Y are both binary vectors (Content Based Filtering | Machine Learning, 2022).

2.1.2 Collaborative-filtering recommender system

Collaborative strategies use the degree of similarity among users to their advantage. The first step in this approach is identifying a cohort or assemblage of persons (referred to as the "neighborhood of A") whose preferences, dislikes, and likes closely to align with those of user A as shown in Figure 3. Subsequently, User A is provided with suggestions for things that are highly favored by most individuals in X. The efficacy of a collaborative algorithm is assessed based on its ability to accurately identify the target user's vicinity. Conventional collaborative filtering systems encounter challenges in terms of privacy and cold start issues due to their reliance on user data sharing. However, it is worth noting that collaborative filtering algorithms do not need any knowledge of item attributes. Moreover, this strategy has the potential to augment the user's current interests via the exploration of new products. There are two distinct categories of collaborative techniques: those that rely on models and those that are based on memories. Memory-based collaborative approaches provide suggestions for new goods by analyzing the preferences of individual users. They use the utility matrix directly in their prediction approach. The first step in this method is to construct a model. The utility matrix serves as the input to the function that represents the model (Dutta & Roy, 2022). The utility matrix is represented by the model f. A function is used to create ten suggestions, using the model and user profile as input. Our suggestions are only accessible to those whose profiles satisfy the criteria of the utility matrix. Due to the need for recalculating the similarity matrix and including the user profile in the utility matrix, this technique requires significant processing resources.

The recommendation function is defined as a function of the user's profile and a given model. In this function, the utility matrix represents the user's profile. Item-based collaborative filtering and user-based collaborative filtering are two subcategories of memory-based collaborative approaches. The user-based approach determines the rating of a recently submitted item by identifying other users within the user's neighborhood who have previously rated the same thing. If the item garners favorable ratings from the user's local community, it is advisable to them.

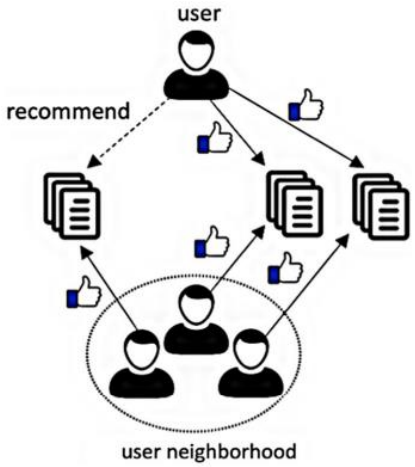

Figure 3- User-based collaborative filtering. Taken from (Dutta & Roy, 2022).

Using the item-based method, all related things that the user has already evaluated are combined to form an item-neighborhood. Then, as seen in Figure 4, the weighted average of all ratings found in an item-neighborhood comparable to the new item is used to estimate the user's rating for that different item.

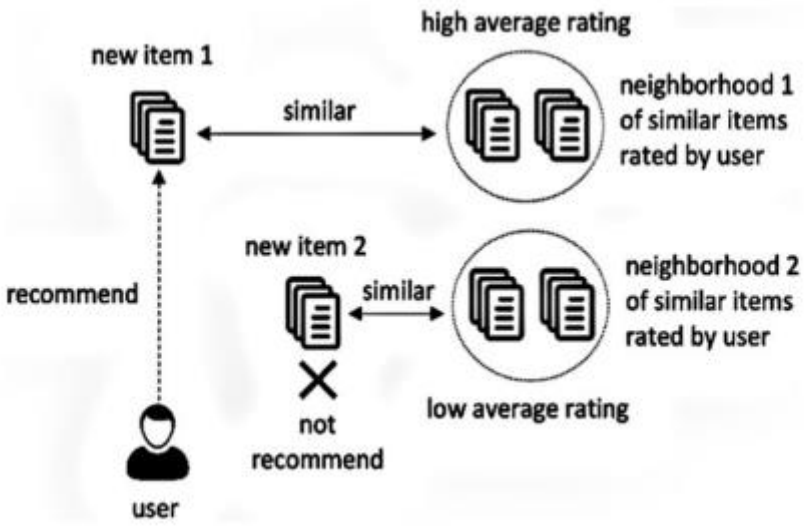

Figure 4 - Item-based collaborative filtering. Taken from (Dutta & Roy, 2022).

2.2 Machine learning methods

2.2.1 Machine Learning Overview

Machine Learning is used to solve various problems that require learning on the part of the machine.

A learning problem has three features:

- Task classes (The task to be learned)
- Performance measures to be improved.
- The process of gaining experience (Alzubi et al., 2018)

Below is illustrated in Figure 5 the machine learning process or workflow in general.:

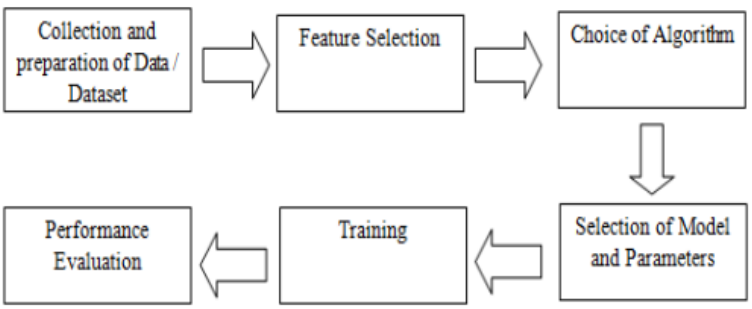

Figure 5 - Components of a Generic ML model. Taken from (Alzubi et al., 2018).

2.2.2 Supervised Learning

Supervised learning is a facet of machine learning that involves the process of acquiring knowledge from labelled training data, which comprises sets of example input-output pairs. The objective is to learn a function that effectively maps input data to corresponding outputs (Mahesh, 2019). In this paradigm, supervised machine learning algorithms require external guidance. The input dataset is typically partitioned into training and test sets, with the training set containing the output variable to be predicted or classified. Various algorithms are employed in this context.

2.2.3 Unsupervised Learning

Unsupervised learning differs from supervised learning as it operates without predefined correct answers or a teacher. In this context, algorithms autonomously explore and unveil interesting structures within the data. Unsupervised learning algorithms extract key features from the data, and when presented with new data, they leverage the learned features to classify or recognize the data's characteristics. This approach is particularly employed for tasks such as clustering and feature reduction (Mahesh, 2019).

2.2.3.1 K-Means Clustering

One of the most straightforward unsupervised learning algorithms for resolving the well-known clustering problem is K-means (Mahesh, 2019). The process uses an easy-to-understand method to categorize the data set via a specific number of clusters. Determining k centers—one for each cluster—is the basic notion. These centers need to be positioned cleverly since different locations yield varied outcomes. Placing them as far apart as feasible is hence the preferred option, as illustrated in Figure 6.

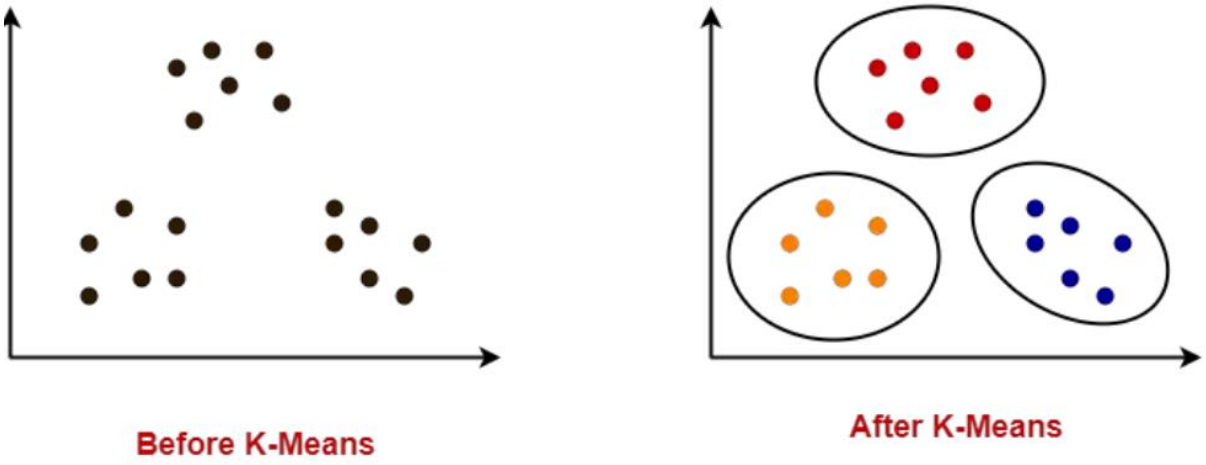

Figure 6 - K-Means Visualisation.Taken from (Marchello, Setiawan, & Gaol, 2019)

For instance (Yoloan et al.2023) implemented k-means clustering to analyze transaction patterns and seasonal correlations in an online retail shop. This study demonstrated the effectiveness of k-means clustering in segmenting customers based on their purchasing behavior, allowing the retail shop to develop strategic sales plans tailored to different customer segments. Another practical example is presented in the work of Rochman et al. (2020) in which k-means clustering to classify tourist destinations based on the number of visitors. This method grouped attractions into categories of 'quiet', 'moderate', and 'crowded', enabling targeted marketing and resource allocation to optimize tourist distribution and enhance visitor experiences

2.3 Sentiment analysis

In some cases, especially when using computers, people's feelings are understood using emotion detection programs. An important part of human-computer interaction is recognizing users' emotions. Computers that can sense emotions could create better connections with people (Alslaity & Orji, 2024). Also, finding emotions helps make systems that can change based on how users feel. Studies in this area have grown because computers can now understand big data and find emotions in it. With the growing attention, it is important to research this field and give a detailed view of the situation today.

For instance, A study by Ghasemaghaei et al. (2016) conducted sentiment analysis on consumer reviews for auto, home, and life insurance services. The research aimed to understand consumers' attitudes toward different types of insurance and predict review ratings based on sentiments. The findings revealed that, since 2013, consumers generally exhibit more negative sentiments toward insurance services. Additionally, the sentiment analysis demonstrated a high accuracy in predicting consumer review ratings, highlighting significant differences between positive, neutral, and negative reviews. This study underscores the importance of sentiment analysis in evaluating consumer attitudes and its potential application in predicting review outcomes.

2.4 Web Scraping

2.4.1 Overview of Web Scraping

Web scraping refers to the process of programmatically extracting information from websites. This technique allows for the collection of vast amounts of data that would otherwise be time-consuming and impractical to gather manually. The basic steps of web scraping include sending an HTTP request to a web page, parsing the HTML content, and extracting the required information using various tools and libraries such as BeautifulSoup, Selenium, and Scrapy (Mitchell, R, 2018).

Using «web scraping," researchers can gather data from several websites such as Google Maps in our case, get customer reviews, and compile it into a single CSV

or database, making the process of analyzing data easier. Web scraping can be a useful technique in machine learning research to gather large volumes of data from many sources, which can subsequently be used to train and test machine learning models. Furthermore, for the following reasons, web scraping is very important in machine learning research:

- Data can be gathered via online scraping from a variety of sources, such as news websites, social networking platforms, and e-commerce websites. This may result in machine learning models that are more detailed and broadly applicable.
- Web scraping can assist researchers in adding new data to already-existing datasets, enhancing the precision and resilience of machine learning models.
- By gathering information on rival companies and their goods, web scraping may be utilized for competitive analysis (Sirisuriya, 2023). This can assist companies in creating machine learning models that can forecast customer behavior, spot market trends, and help them make strategic decisions.

In the fundamental aspects of web scraping the objective is to streamline the process of collecting internet information and converting it into a coherent dataset. In Figure 7, we can see how using three distinct technologies, namely Selenium, WebDriver, and XPath, can help create a scraper that simplifies the process of gathering data from prominent websites (Han & Anderson, 2021).

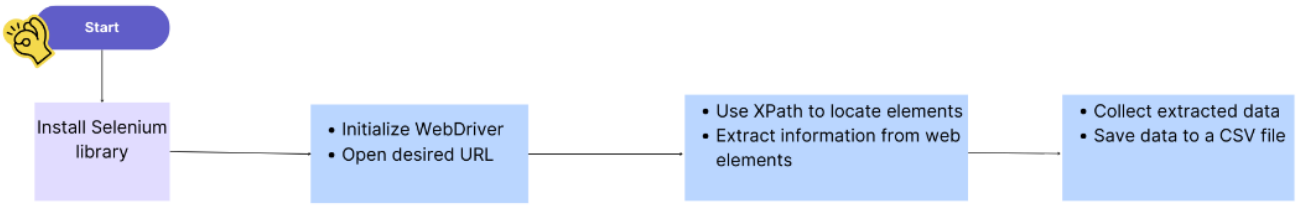

Figure 7 - Workflow for Web Data Extraction Using Selenium

2.4.2 Tools and techniques

Selenium: Selenium is a tool that uses a collection of open-source APIs called WebDriver to automate the testing of web applications (Francis Academic Press, 2023). It provides a robust framework for navigating web pages, interacting with web elements, and extracting data. Selenium is particularly useful for scraping dynamic web pages that require JavaScript execution

BeautifulSoup: Beautiful Soup is a Python library used for parsing HTML and XML documents. It provides simple methods for navigating and searching the parse tree, making it easier to extract the desired data from web pages (Richardson, 2024).

2.4.3 Ethical Considerations

Ethical considerations are paramount when conducting web scraping, particularly with sensitive data such as Google Maps reviews (Han & Anderson, 2021). Privacy must be preserved, ensuring that individuals' personal information is protected. For instance, when scraping Google Maps reviews, it is crucial to handle the data responsibly, anonymizing user information, and adhering to data privacy laws. Using this data with a consciousness of its potential impact on privacy and maintaining respect for the terms of service of the platform is essential. This approach helps avoid legal issues and fosters trust and integrity in the research or data products being developed.

3. METHODOLOGY

The methodology section will explain how we developed a tool recommending the best repair shop to customers. We have further elaborated on the methods mentioned above in the CRISP-DM framework. Next, we present the datasets and perform exploratory analysis to uncover critical insights. We then move to the discussion of the data engineering process, where we preprocess the data for modeling and present the model process.

3.1 Methodology Using CRISP-DM

We used the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is today the most widely used approach for carrying out data mining and data science projects (Kannengiesser & Gero, 2023), to ensure the development of a systematic and effective car bodyshop recommendation system was done correctly. CRISP-DM has provided a kind of general framework from understanding business objectives to final model deployment. The detailed steps are presented as follows:

Business Understanding:

- **Objective**: The goal was to recommend the best repair shop to the customer. This recommendation system is designed to optimize customer satisfaction and the cost of repairs. The following were sub-goals: optimize the precision of recommendations for repair shops, minimize total repair cost, and maximize customer satisfaction by leading them to the right shop.

Data Understanding:

- **Exploration**: Detailed exploration of the dataset to understand its structure, what it contains, and the relationship between variables. The collected data was from various sources.

Data Preparation:

- **Cleaning**: Data cleaning includes dealing with missing data, finding duplicates, and filtering out junk observations such as those regarding total loss or fraud claims.
- **Transformation**: Necessary transformations were made to make data formats consistent to maintain the consistency of the format in all datasets.
- **Integration**: It was the process through which datasets were integrated based on shared identifiers to result in one dataset that could be implemented for modeling. This ensured proper and consistent data was ready for other analyses.

Modeling:

- **Clustering Algorithms**: Complex clustering algorithms are set up to ensure the proper grouping of claims and bodyshops independently. This mainly involves techniques like K-Means clustering, which finds patterns and data relationships.
- **Feature selection:** Features were selected that play a vital role in clustering results. Such features included operational metrics, financial performance, and customer satisfaction scores.

Evaluation:

- **Performance Metrics**: Clustering models were used to measure the performance with proper metrics that could describe the quality and effectiveness of the clusters. This process assures that the models developed will truly reflect the data and ensure that inferences made are meaningful.
- **Validation**: The clusters were validated against known benchmarks and expert evaluations so that their relevance and correctness was confirmed.

3.2 Data Collection and Exploratory Analysis

3.2.1 Dataset Description:

The dataset comprises several distinct datasets, each offering unique insights into different aspects of the insurance and repair process. The data was collected from various sources such as the data warehouse Greenplum, the business unit, and Google Maps scraped reviews. It is highly confidential and secure. Here is a detailed description of each dataset:

The **Appraisals Dataset** (presented in Table 1) contains information about appraisals conducted for vehicle repairs, including repair costs, types of repairs, and appraisal dates starting in 2021. This dataset is crucial for understanding the financial and temporal aspects of vehicle repair processes.

Table 1 - Appraisals Dataset

Column	Type	Description	Key
numero_ocorrencia	Id	Occurrence id	Yes
numero_sinistro	Id	Claim id	Yes
numero_companhia	Id	Company id	Yes
result_id	Id	Result id	Yes

Column	Type	Description	Key
expertise_id	Id	Expertise id	Yes
tipo_resultado_servico	category	Type of result	No
sub_tipo_servico	category	Sub-type of service	No
desc_sub_tipo_servico	category	Sub-type of service - description	No
data_criacao_rs	date	Date of result creation	No
expertise_type	category	Type of expertise	No
bodyshop_id	id	Bodyshop id	No
valor_total_sem_iva_orcamento	numeric	Total repair cost before taxes, retention, and coverage deductible (€)	No
qtd_mao_de_obra_em_segundos_orc_per_resumo	numeric	Labour time (in seconds)	No
valor_pecas_orc_per_resumo	numeric	Parts cost (€)	No

The **Bodyshops Dataset** (presented in Table 2) provides details about various repair facilities, including bodyshop ID, location, type of bodyshop, and operational status. This dataset is essential for analyzing the operational characteristics and geographical distribution of bodyshops.

Table 2 - Bodyshops Dataset

Column	Type	Description	Key
bodyshop_id	id	Bodyshop id	Yes
nome_prestador_oficina	category	Name of bodyshop	No
descricao_situacao_prestador_oficina	category	Status of bodyshop (active or not)	No
zona_local_prestacao_oficina	id	Region of bodyshop	Yes
local_prestacao_oficina	id	Location of bodyshop	Yes
ind_convencionada	category	Indicator if bodyshop is (in-network)	No
type_bodyshop	category	Type of bodyshop	No
bodyshop_district	category	Bodyshop district	No
bodyshop_municipality	category	Bodyshop municipality	No
bodyshop_address	category	Bodyshop address	No
CAPITAL	numeric	Capital	No
DATACONST	date	Date of establishment	No
VALOR_SCORE	category	Rank of the probability of company defaulting in the next 12 months	No
N_EMPREGADOS	numeric	Number of employees	No
TOTAL_ACTIVO_LIQUIDO_2022	numeric	Total net assets in 2022	No
RESLIQ_2022	numeric	Net results in 2022	No

The third dataset, detailed in Table 3, is the **Claims Dataset**. This dataset includes comprehensive information about insurance claims, such as claim types, coverage details, claim creation dates, and vehicle categories.

Table 3 - Claims Dataset

Column	Type	Description	Key
numero_ocorrencia	id	Occurrence id	Yes
numero_sinistro	id	Claim id	Yes
numero_cliente	id	Client id	Yes
numero_golden_record	id	Client golden record id (one person can have only one golden record but multiple client numbers)	Yes
numero_companhia	id	Company id	Yes
data_ocorrencia	id	Date of damage event	No
data_abertura_ocorrencia	date	Occurrence opening date	No
codigo_cobertura	category	Claim coverage	No
flag_investigacao	flag	Indicator if occurrence was investigated/inquired (clarifications, missing documents, etc.)	No
ind_total_loss	flag	Indicator of total loss	No
ind_fraud	flag	Indicator of fraudulent occurrence	No
pos1	flag	Indicator of left side Hit Point - Front	No
pos2	flag	Indicator of right side Hit Point - Front	No
pos3	flag	Indicator of left side Hit Point - Center	No
pos4	flag	Indicator of right side Hit Point - Center	No
pos5	flag	Indicator of left side Hit Point	No

Column	Type	Description	Key
pos6	flag	Indicator of right side Hit Point	No
pos7_dir	flag	Indicator of rear Hit Point - Right	No
pos7_esq	flag	Indicator of rear Hit Point - Left	No

The fourth dataset, detailed in Table 4, is the **Vehicles Dataset**. This dataset offers insights into vehicle characteristics, such as vehicle categories, which are essential for understanding repair requirements and costs.

Table 4 - Vehicles Dataset

Column	Type	Description	Key
numero_ocorrencia	id	Occurrence id	Yes
numero_sinistro	Id	Claim id	Yes
numero_companhia	Id	Company id	Yes
identificador_objecto_sin	Id	Object id	Yes
matricula_objecto_sin	Id	License plate	Yes
ano_veiculo_sin	numeric	Year of vehicle manufacturing	No
tipo_relacao_objecto_sin	category	Relation with claim	No
numero_chassis_sin	id	VIN	No
marca_veiculo	category	Vehicle brand	No
modelo_veiculo	category	Vehicle model	No
categoria_veiculo	categoria_veiculo	Vehicle category	No

The fifth dataset, as detailed in Table 5, is the **Spare Parts Dataset**. This dataset contains detailed information about spare parts involved in the repairs, including the cost amount and descriptions.

Table 5 - Spare Parts Dataset

Column	Type	Description	Key
numero_sinistro	Id	Claim id	Yes
numero_companhia	Id	Company id	Yes
expertise_id	Id	Expertise id	Yes
referencia_peca_orc_per_det_pecas	Id	Id of bodypart	Yes
referencia_peca_escolhida_orc_per_det_pecas	Id	Id of bodypart chosen	No
valor_pecas_orc_per_det_pecas	numeric	Cost amount (€)	No
descricao_orc_per_det_pecas	category	Description of body part	No

The sixth dataset, as detailed in Table 6, is the **Painting Costs Dataset**. This dataset contains detailed information about painting parts involved in the repairs, including labor quantity, labor cost, material cost, and descriptions.

Table 6 - Painting Costs Dataset

Column	Type	Description	Key
numero_sinistro	Id	Claim id	Yes
numero_companhia	Id	Company id	Yes
expertise_id	Id	Expertise id	Yes
referencia_peca_orc_per_det_pintura	Id	Painting part reference	Yes
valor_mao_de_obra_pintura_orc_per_det_pintura	numeric	Painting labor cost	No

Column	Type	Description	Key
qtd_mao_de_obra_pintura_em_segundos_orc_per_det_pintura	numeric	Painting labor quantity in seconds	No
valor_do_material_pintura_orc_per_det_pintura	numeric	Painting material cost	No

In addition to the above datasets, a separate dataset was created by scraping data from Google Maps. This dataset includes reviews of bodyshops and their respective ratings, providing valuable insights for sentiment analysis and final clustering. The Google Maps Reviews Dataset will be discussed in detail in the data exploration and sentiment analysis sections, where it will be used to rank bodyshops and perform sentiment analysis to aid in the final clustering of bodyshops.

3.2.2 Data Exploration:

An exploratory data analysis, or EDA, is a crucial stage in figuring out the underlying patterns, connections, and insights in the dataset. This stage entails a detailed analysis of the data to find significant patterns, spot anomalies, and develop a full grasp of the variables and how they interact. Through the application of diverse statistical and visual aids, EDA establishes a framework for further modelling and analysis, guaranteeing that the data is comprehensible and ready for increasingly complex analytical assignments.

We will examine the main conclusions from our exploratory study in this section, paying particular attention to elements like customer reviews, bodyshop performance, and repair costs. The objective is to derive practical insights that will guide our modelling choices and improve the repair shop recommendation system's overall efficacy. For instance, the **Average Repair Cost Analysis** aimed to compare average repair costs between repairs conducted within and outside the insurance network. The method involved calculating the mean repair costs for both in-network and out-of-network repairs. As seen in Figure 8, the analysis revealed that in-network repairs tend to be less cost-effective compared to out-of-network repairs. This suggests potential benefits for both insurers and customers when opting for out-of-network repair facilities in some cases. By systematically exploring these areas, we aim to build a robust understanding of the data, laying the groundwork for developing effective clustering models and ultimately enhancing the repair shop

recommendation system. The following sections detail the methodologies and findings for each of these objectives.

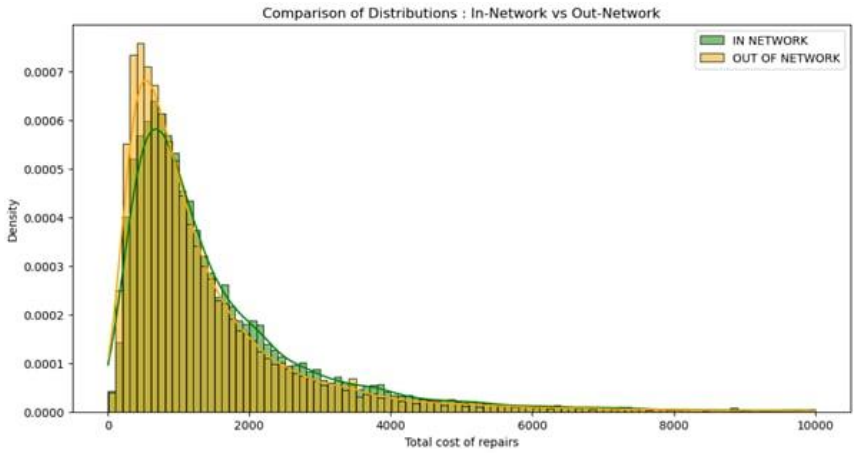

Figure 8 - Comparison of distributions: In network vs out of network

Brand dealers vs. Multi Brand vs. Collision Centers, are the three bodyshop types in insurance Brand dealers are bodyshops dealing with certain brands, Multi brands are generic bodyshops that accept in most cases all brands, and collision centers are few bodyshops strictly linked to the insurance and holds its brand among other advantages mainly in services. This analysis focused on comparing the average repair costs among branded repair shops, non-branded repair shops, and collision centers. The method grouped the repair shops into three categories and calculated the average repair costs for each category. As shown in Figure 9, collision centers emerged as the most cost-effective repair solution, likely due to their specialized equipment and streamlined processes. Branded repair shops were generally more expensive, while non-branded repair shops offered a middle ground.

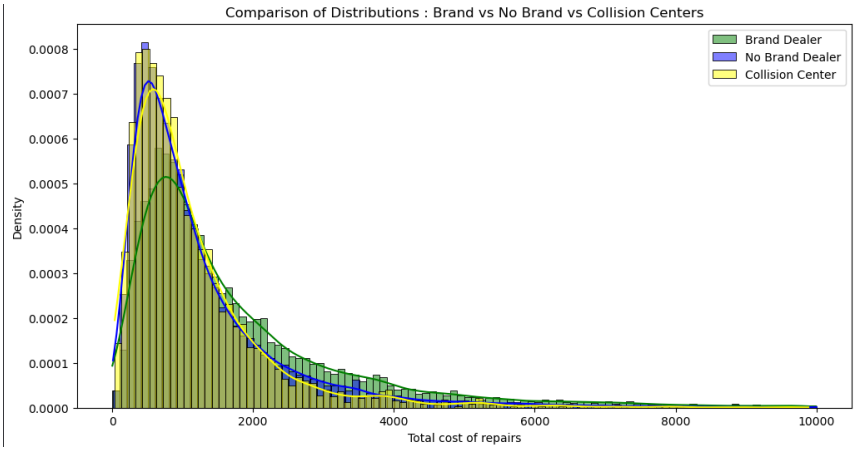

Figure 9 - Comparison of distributions: In network vs out of network

The **Median Repair Cost Analysis by Brand and Bodyshop Type** further explores the cost differences across vehicle brands and bodyshop types. The objective was to analyze the median repair costs for brand and non-brand dealers across different vehicle brands. The method calculated the median cost difference between in-network and out-of-network repairs for each brand and bodyshop type. As seen in Figure 10, significant variations were observed across brands. For example, Ford and Renault's repairs were notably more expensive at brand dealers, while brands like Opel and Fiat showed cost savings compared to non-brand dealers.

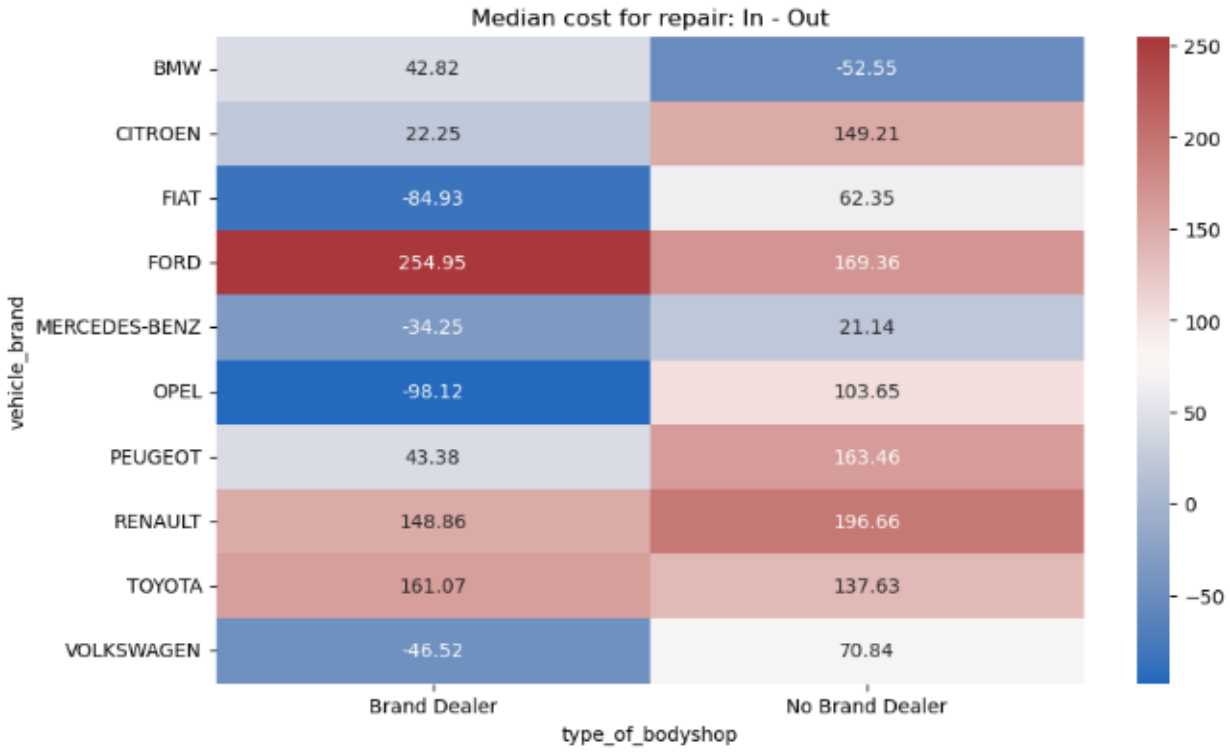

Figure 10 - Median cost of repair: In /Out

The **Percentage of Claims per Bodyshop Type** analysis aimed to examine the distribution of claims across various bodyshop types relative to vehicle age. The method involved computing the percentage of claims handled by each bodyshop type and examining the correlation with the age of the vehicles. As depicted in Figure 11, customer preferences for bodyshop types varied based on factors such as brand affiliation and vehicle age, with newer vehicles more likely to be repaired at branded shops, while older vehicles were often serviced at non-branded or collision centers. This analysis highlighted significant patterns in repair choices, revealing that older vehicles tend to be repaired at generic out-of-network bodyshops, which handle a higher percentage of claims as vehicle age increases.

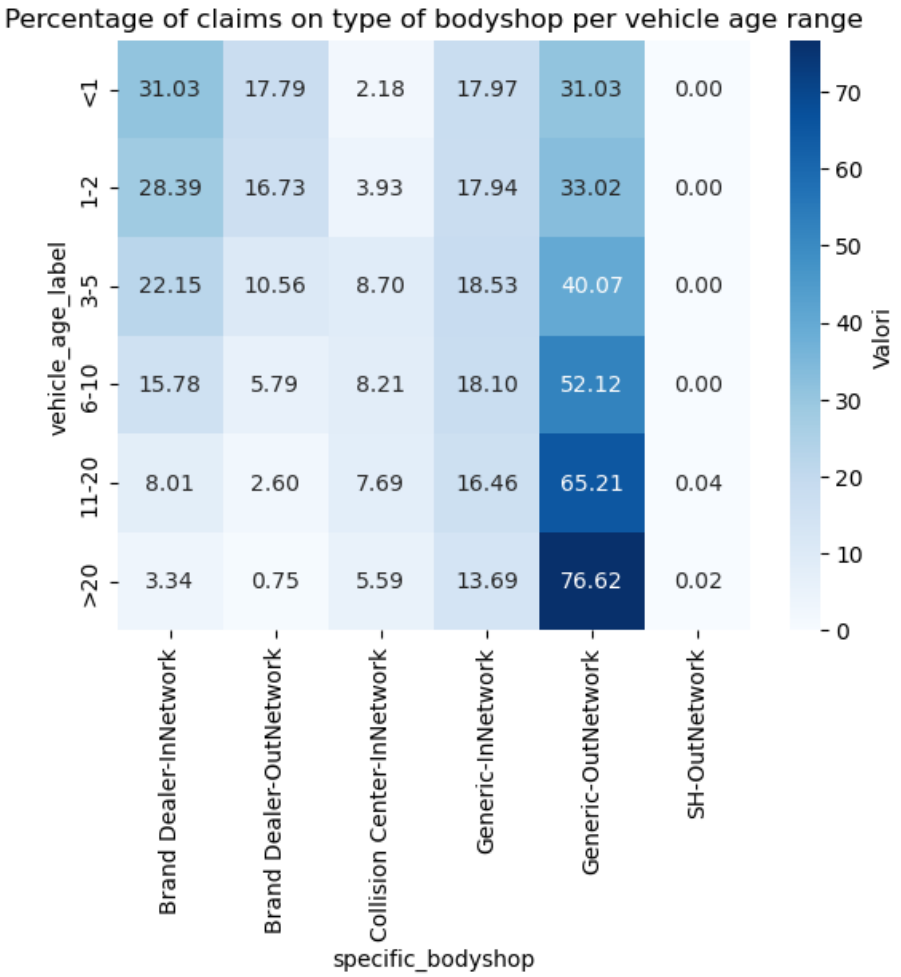

Figure 11 - Claims per type of bodyshop per vehicle age

By systematically exploring these areas, we aim to build a robust understanding of the data, laying the groundwork for developing effective clustering models and ultimately enhancing the repair shop recommendation system. The following sections detail the methodologies and findings for each of these objectives.

3.3 Data Engineering

3.3.1 Data Transformation

The data underwent a series of merging operations to combine information from all the different sources. This included merging appraisals with bodyshops, claims, and policies, ensuring that all relevant information was integrated. Filters were then applied to keep only appraisals in the scope. For instance, electric vehicles, specific vehicle types, and certain coverages were removed. Also, only appraisals from 2023 were kept as training data for ranking.The dataset was further filtered to remove any records with vehicle manufacture years before 1900. Additional columns were created, such as a unique bodyshop ID, to facilitate the identification of distinct bodyshops.

3.3.2 Feature Engineering

The code creates a new column called hitpoint_agg, which groups various specific collision points into broader categories like "front," "back," and "lateral." This new column simplifies the data by consolidating detailed collision points into more general ones, making it easier to analyze repair costs by these broader categories.

After that, the total repair costs for each unique collision point were calculated, which determines the percentage contribution of each type of repair cost to the total. It stores these average percentage contributions in a dictionary, organized by collision point and repair cost type. This distribution will help later rank bodyshops based on hit point.

The dataset was analyzed to rank bodyshops based on various repair costs. New columns for labor costs, unique bodyshop IDs, and discount percentages. The data is filtered to include only 2023 records, remove negative and extreme values, and ensure repairs involve at least one hour of labor. Bodyshops with fewer than five claims historically are excluded. Then discounted labor costs per hour were calculated, aggregated by bodyshop, and computed interquartile means to score bodyshops based on labor and spare parts costs. For painting costs, it calculates the interquartile mean for painting material costs per bodyshop and scores them based on their discount percentages relative to the median. These scores are then merged, and a final ranking for bodyshops is computed based on a weighted average of scores, considering the distribution of repair costs by collision point (hit point). This results in detailed and organized rankings of bodyshops, allowing for a nuanced assessment of repair costs by different collision points.

Through this analysis, we created key columns such as average_total_repair_without_tax, which represent the average cost of repairs excluding tax by averaging the total repair costs for each bodyshop. We also derived average_days_of_repair, calculated by averaging the repair durations for all claims handled by the bodyshop to estimate how fast the bodyshop, in general, is, and average_labour_hour_discounted as an indicator for labor cost representing the average cost per hour for labor adjusted for discounts.

3.3.3 Google Maps Reviews Analysis

3.3.3.1 Reviews Scraping

To collect customer reviews for bodyshops, we implemented a web scraping process using Python libraries Selenium and BeautifulSoup. Selenium automated the necessary browser interactions to access the reviews, while BeautifulSoup handled parsing and data extraction. The scraping process began with the initialization of a Selenium WebDriver instance, which opened the Google Maps page of each bodyshop. The WebDriver navigated to the reviews section and scrolled down to load all available reviews, ensuring comprehensive data collection. Using BeautifulSoup, we parsed the HTML content to extract review texts, ratings,

and timestamps, storing this data in a structured format for subsequent analysis. This method ensured we gathered relevant data, capturing the true voice of the customers.

3.3.3.2 Sentiment Analysis

After collecting the reviews, the next step was to analyze the sentiment expressed in them. We employed a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model for this purpose, leveraging its advanced natural language understanding capabilities. The preprocessing phase involved cleaning the review texts by removing duplicates, handling missing values, and standardizing text formats. This process included converting text to lowercase and removing punctuation and stop words to ensure consistency.

For the sentiment analysis, we utilized a pre-trained BERT model that was fine-tuned on a labeled dataset of customer reviews, adapting it to our specific task. This model predicted sentiment scores for each review, assigning a score from 1 (very negative) to 5 (very positive). The fine-tuned model provided accurate sentiment predictions, allowing us to quantify customer satisfaction effectively.

The sentiment scores were then integrated into our clustering algorithm as additional features. This integration provided a more nuanced understanding of customer satisfaction, enabling us to identify patterns and insights that aligned with operational metrics and customer experiences. The results from the sentiment analysis were crucial in enhancing the accuracy and relevance of the final bodyshop recommendations, ensuring that customer satisfaction was a key component of the decision-making process.

3.4 Clustering

The clustering analysis was conducted using the K-Means algorithm to group bodyshops based on several key metrics. The process involved several critical steps, including data cleaning, feature selection, and determining the optimal number of clusters.

3.4.1 Data Cleaning

The initial dataset contained various operational and financial metrics for each bodyshop. To ensure data quality, we took several steps to preprocess the data.

We handled missing values by either imputing the median like in the case of number of employees and capital or removing them to keep the dataset consistent. All features were standardized to have a mean of 0 and a standard deviation of 1, which ensured that each feature had equal weight in the clustering process. Outliers were identified and removed to prevent them from skewing the clustering results. For example, bodyshops with fewer than 5 reviews were excluded to ensure the analysis was based on sufficient customer feedback.

3.4.2 Feature Selection

The features selected for clustering included both operational and financial metrics, as well as customer satisfaction scores derived from sentiment analysis. The key features used were the combined customer satisfaction score, total number of ratings, number of employees, total net assets in 2022, total parts count, percentage of non-original parts, scores for painting material cost and spare parts material cost, average total repair cost without tax, rank back, date of establishment, risk score, net profit in 2022, average days of repair, average labor hour discounted, number of claims, and an indicator of whether the bodyshop is in-network.

To ensure the relevance and effectiveness of these features in the clustering process, we conducted a correlation analysis. This analysis helped us understand the relationships between different features and identify any highly correlated features. To keep only the features, we wanted to explore and avoid highly correlated features that can affect the clustering. As seen in Figure 12, the correlation matrix provided insights into these relationships and guided our feature selection process.

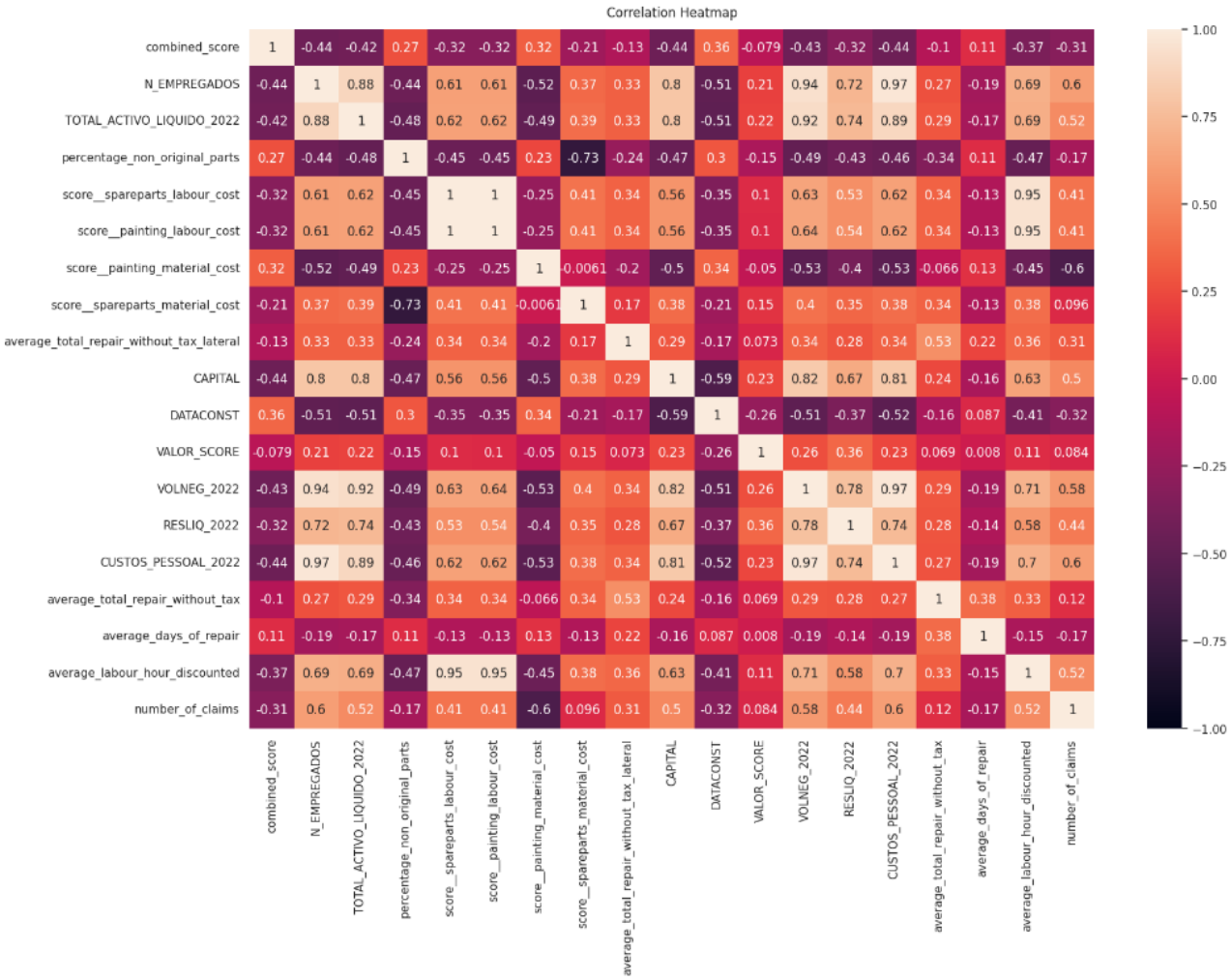

Figure 12 - Correlation Heatmap of Bodyshop Dataset Variables

4. RESULTS AND DISCUSSION

4.1 Results

To determine the optimal number of clusters, we employed both the Elbow Method and the Silhouette Score Method (Öztürk & Demirel, 2023). These methods provide complementary perspectives on evaluating clustering performance. The Elbow Method involved running the K-Means algorithm for a range of cluster numbers and plotting the Within-Cluster Sum of Squares (WCSS) for each (Öztürk & Demirel, 2023). The point where the WCSS started to level off, known as the elbow point, indicates the optimal number of clusters by balancing model complexity and variance reduction. As depicted in Figure 13, the elbow point suggested that four clusters were optimal. In addition to the Elbow Method, we also used the Silhouette Score Method, which measures how similar each data point is to its cluster compared to other clusters. A higher silhouette score indicates better-defined clusters. As illustrated in Figure 14, the silhouette scores for different numbers of clusters showed a significant drop beyond four clusters, supporting the decision to use four clusters.

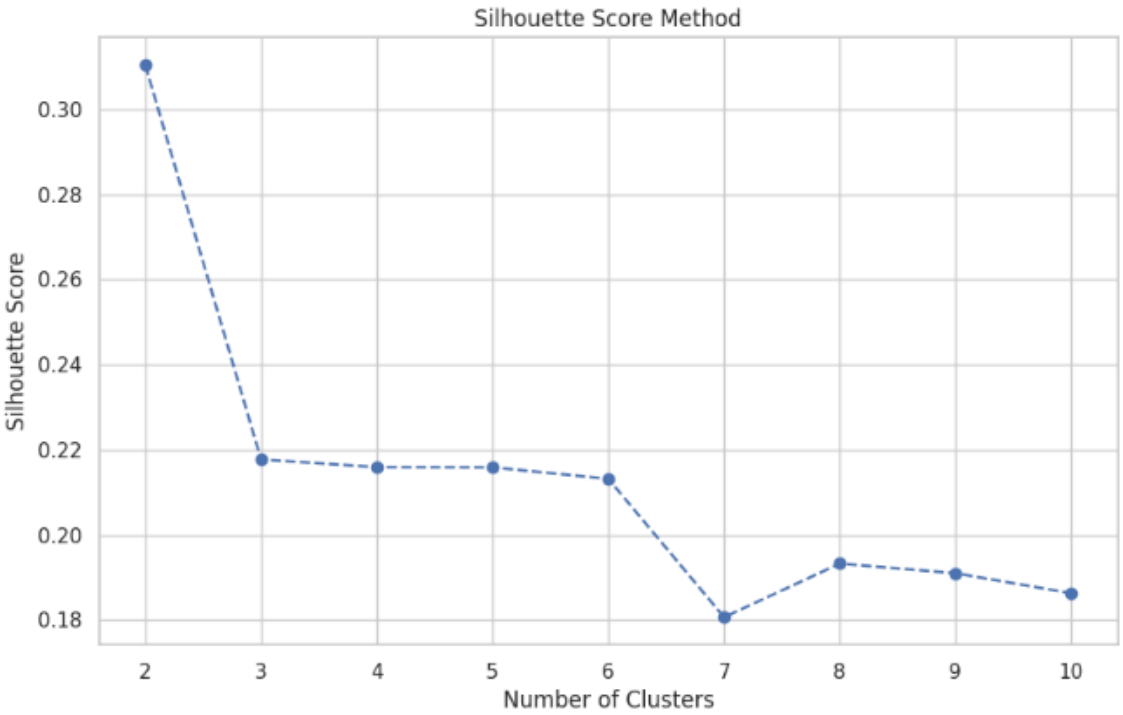

Figure 13 - Silhouette Score Method for Determining Optimal Number of Clusters

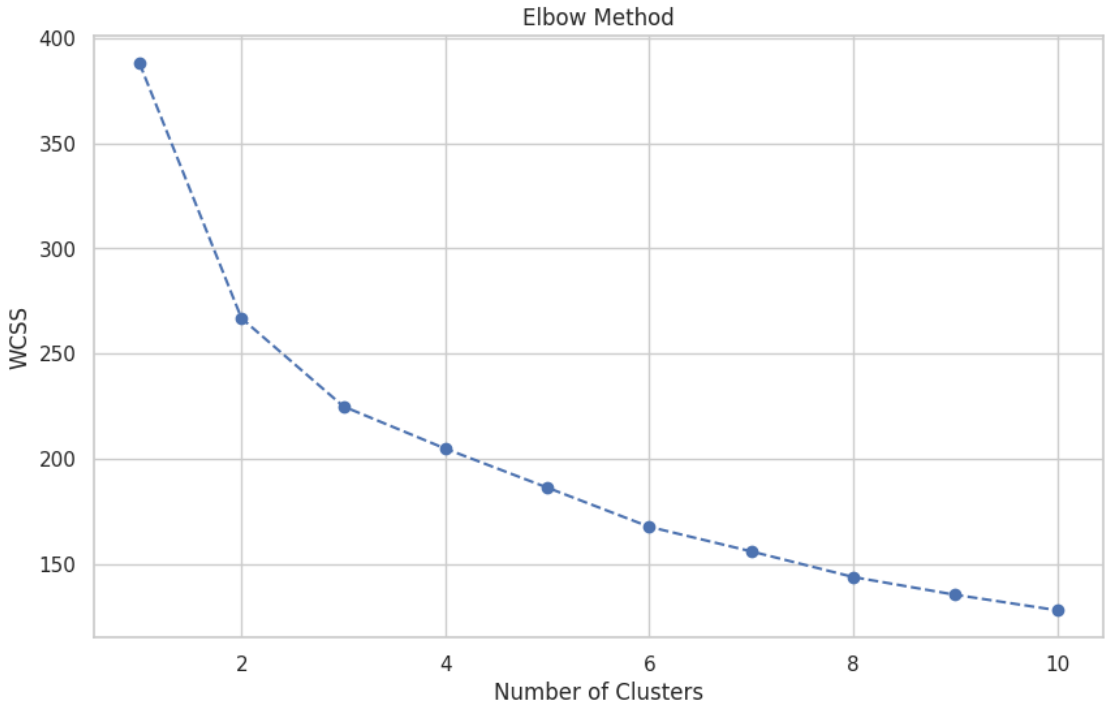

Figure 14 - Elbow Method for Determining Optimal Number of Clusters

Based on the analysis using these two methods, we chose to use four clusters for our clustering analysis. The initial decision was to choose 2 clusters .However,the choice of four clusters balances the trade-off between complexity and the quality of clustering. Once the optimal number of clusters was identified, the K-Means algorithm was applied to partition the bodyshops into distinct groups. Each cluster represented bodyshops with similar operational and financial characteristics, providing insights into the different types of bodyshops and helping to identify clusters with high customer satisfaction and cost-effective operations

In this section, we present the findings from the clustering analysis of bodyshops. The bodyshops were grouped into four distinct clusters, each with unique characteristics.

4.1.1 Cluster Descriptions

Before diving into the detailed analysis of each cluster, we provide an overview of the composition and network status of the clusters using three key tables. First, we examine the types of bodyshops within each cluster, as shown in Table 7

Table 7 - Cluster Composition by Type of Bodyshop

Type of Bodyshop	Collision Centers	Brand Dealers	Multi brand
Cluster 0	10	22	449

Type of Bodyshop	Collision Centers	Brand Dealers	Multi brand
Cluster 1	0	0	399
Cluster 2	0	20	6
Cluster 3	1	149	99

Next, we look at the network status of the bodyshops, detailing how many are in-network versus out-of-network, as presented in Table 8.

Table 8 - Cluster Composition by Network Status

In-Network Status	Out of Network	In Network
Cluster 0	398	83
Cluster 1	313	86
Cluster 2	2	24
Cluster 3	127	122

After that, the parallel coordinates plot in Figure 15 was performed to provide a better visualization of the clustering analysis across the numerical attributes. Each line represents one of the four clusters (0, 1, 2, and 3), showcasing how they differ in each feature. The values of these features have been normalized between 0 and 1 to facilitate comparison. Notable patterns include the distinct behavior of cluster 2 with higher values in features like total ratings and TOTAL_ACTIVO_LIQUIDO_2022, while clusters 0 and 1 show significant variation in features like percentage non original parts and score spare parts material cost. This plot aids in understanding the unique characteristics and similarities among the clusters.

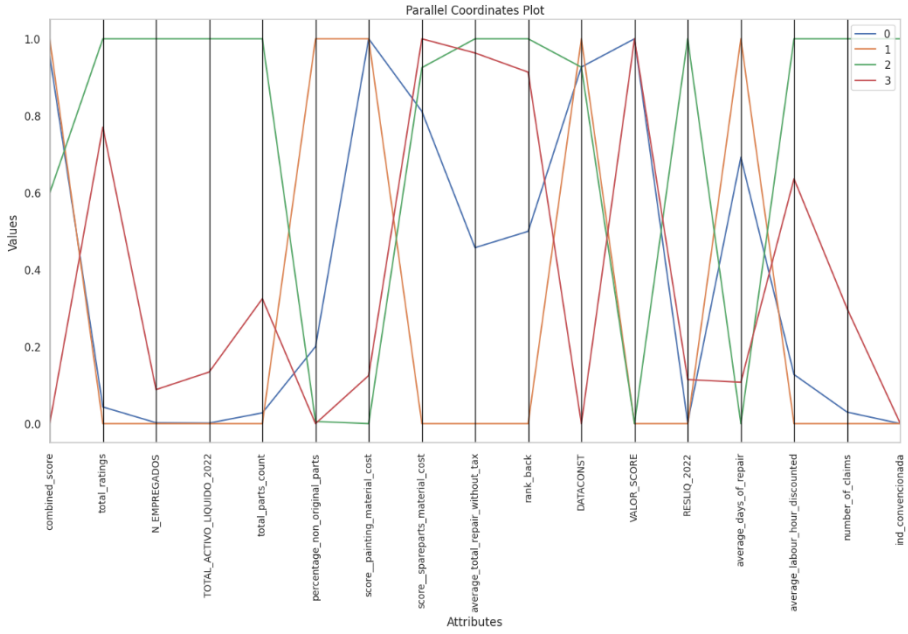

Figure 15 - Parallel Coordinates Plot for Clustering Analysis

Finally, we summarize the key characteristics of each cluster, highlighting metrics such as combined score, total ratings, number of employees, and capital, as shown in Table 9.

Table 9 - Cluster Characteristics Summary

Metric	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Combined Score	4.5	4.51	4.29	4.00
Total Ratings	65	47	272	245
Number of Employees	17	8	770	118
Total Parts Count	93	57	452	204
Percentage of Non-Original Parts	7.2	26.9	1.6	0.9
Painting Material Cost Score	-2.22	-2.44	-8.23	0.06
Spare Parts Material Cost Score	-9.24	-25.87	-6.58	0.78
Average Repair Cost	€1,396.27	€1,127.00	€1,680.67	€1,562.42
Rank	630	211	1069	1000
Date of Establishment	1990	1998	1997	1983
Average Labour Hour	34.87	31.25	57.82	49.77
Number of Claims	28	18	116	46

Following these tables, we dive into a detailed analysis of each cluster to highlight their unique characteristics and implications for optimizing the selection of bodyshops.

Cluster 0: "Best Collision and Multibrand Dealers" Cluster 0 includes a mix of collision centers, multi-brand bodyshops, and brand dealers. These bodyshops are highly rated, making them the best collision centers and multibrand options. When recommending bodyshops, those in this cluster should be prioritized based on their ranking. Although this

cluster is the second best in terms of average cost, the variety of bodyshops within it justifies this. The mix of high-quality collision centers and dealers explains the slightly higher average costs, which is understandable considering the service quality they offer.

Cluster 1: "Cost-Effective Generics" Cluster 1 comprises exclusively generic bodyshops. These bodyshops have excellent review scores and offer the best average costs. However, their capacity might be an issue. Therefore, while these bodyshops should be considered for recommendations due to their cost efficiency, they should be used with caution to avoid capacity-related problems. Their lower capacity necessitates careful management to ensure they are not overwhelmed, but their cost-effectiveness makes them a valuable option.

Cluster 2: "Frequent Dealers with Higher Costs" Cluster 2 mainly includes brand dealers and generic bodyshops that we frequently deal with, as evidenced by the high number of claims. Despite their familiarity, these bodyshops are not cost-effective, with a high average cost. The mix of bodyshops sent to this cluster is a contributing factor. Reducing the number of claims directed to these bodyshops could help lower the total average cost. This cluster should be monitored closely, and strategies should be implemented to optimize the claim distribution and manage the mix more effectively.

Cluster 3: "High-Cost, Low-Rated" Cluster 3 is characterized by higher costs and lower ratings, making these bodyshops less favorable. They should generally be avoided to optimize network performance. However, in-network bodyshops within this cluster should be revisited and potentially replaced with willing bodyshops from Cluster 1. This swap could enhance the overall network by incorporating cost-effective and higher-rated options from Cluster 1.

4.2 Discussion

The clustering analysis revealed significant insights into the operational and financial characteristics of different bodyshops. Each cluster demonstrates unique attributes that can help insurance companies optimize their selection of bodyshops for repairs.

Cluster 0: "Best Collision and Multibrand Dealers" consists of a mix of collision centers, multibrand bodyshops, and brand dealers. These bodyshops are highly rated, making them the best options for quality repairs. When recommending bodyshops, those in this cluster should be prioritized based on their ranking. Although this cluster is the second best in terms of average cost, the variety and quality of bodyshops within it justify this.

Cluster 1: "Cost-Effective Generics" includes generic bodyshops with excellent review scores and the best average costs. However, due to their lower capacity, these bodyshops should

be used with caution to avoid overwhelming them. Their cost-effectiveness makes them valuable, but careful management is necessary to ensure they are not overburdened.

Cluster 2: "Frequent Dealers with Higher Costs" comprises brand dealers and generic bodyshops with a high number of claims. Although familiar, these bodyshops are not cost-effective. Reducing the number of claims to these bodyshops could help lower the average cost. This cluster requires close monitoring and strategies to optimize claim distribution and manage the mix more effectively.

Cluster 3: "High-Cost, Low-Rated" is characterized by higher costs and lower ratings, making these bodyshops less favorable. Generally, they should be avoided to optimize network performance. However, in-network bodyshops within this cluster should be revisited and potentially replaced with willing bodyshops from Cluster 1. This swap could enhance the overall network by incorporating cost-effective and higher-rated options from Cluster 1, thereby improving both cost efficiency and customer satisfaction.

5. CONCLUSIONS

This research has demonstrated the successful development and implementation of a data-driven recommender system aimed at optimizing the steering process of car bodyshops for an insurance company in Portugal. By utilizing advanced machine learning techniques within the structured CRISP-DM framework, we were able to identify distinct clusters of bodyshops, each characterized by unique operational and financial attributes that inform strategic decision-making.

The systematic approach began with an extensive data collection process that integrated multiple data sources, including claims data, bodyshop performance metrics, and customer reviews. This comprehensive dataset facilitated a robust exploratory data analysis, uncovering critical insights into the factors influencing repair facility selection. Our methodology emphasized data quality through rigorous cleaning, transformation, and feature engineering processes, ensuring the reliability and accuracy of the clustering models.

The implementation of this recommender system demonstrates significant advancements in balancing cost reduction, customer satisfaction, and operational efficiency within the insurance claims process. The study underscores the transformative potential of integrating data analytics into traditional business operations, providing a blueprint for future innovations in the insurance industry and beyond. The practical insights gained extend the applicability of this approach to other industries where service provider selection is critical, underscoring the versatility and impact of data-driven methodologies.

Future research could build upon this study by incorporating a wider range of data sources, such as telematics data and customer feedback from different platforms, to further enhance the model's accuracy and comprehensiveness. Additionally, exploring advanced machine learning techniques like deep learning and reinforcement learning could yield even more sophisticated and effective recommender systems. Expanding the validation of the model across different geographic regions and insurance companies would also enhance the generalizability of the findings, ensuring broader applicability and relevance.

In conclusion, this research provides a solid foundation for the development and application of data-driven recommender systems within the insurance industry. It highlights the significant benefits of leveraging data analytics for operational optimization, paving the way for more efficient, cost-effective, and customer-centric strategies in managing insurance claims and repair facility selections. The successful implementation of this system not only demonstrates its practical value but also sets a precedent for future advancements in the field, contributing to the ongoing evolution of data-driven business practices.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORK

Despite the promising results, this study has several limitations. Firstly, the analysis is based on data from a single insurance company in Portugal, which may limit the generalizability of the findings. Future research should consider incorporating data from multiple companies and geographic regions to validate and extend the applicability of the model.

A significant limitation is the absence of severity estimation at the time of claim opening, which complicates the prediction of claim costs and the nature of repairs needed. This challenge could be addressed by integrating image analytics in future studies, allowing for a more accurate and immediate assessment of damage through photos submitted at the time of the claim.

Another notable limitation is the lack of customer feedback following appraisals. Such feedback would be invaluable for monitoring the ongoing performance of bodyshops and ensuring continuous improvement in service quality. Incorporating post-appraisal customer feedback could provide deeper insights into customer satisfaction, refining the selection process further and enhancing the recommender system's effectiveness.

Additionally, while the current model primarily focuses on cost and customer satisfaction metrics, it would benefit from considering a broader range of factors. For instance, including environmental impact and the long-term quality of repairs could offer a more holistic view of the bodyshops' performance. This would ensure the system not only optimizes immediate costs and satisfaction but also promotes sustainable and high-quality repair practices.

By addressing these limitations, future research can build on the foundation laid by this study, leading to more comprehensive and generalizable insights that benefit the insurance industry and its customers.

BIBLIOGRAPHICAL REFERENCES

Ahmed, N., & Razak T, A. (2016). IJARCCE An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*. https://doi.org/10.17148/IJARCCE.2016.5277

Al-Hagery, M. (2016). Google Search Filter Using Cosine Similarity Measure to Find All Relevant Documents of a Specific Research Topic. *INTERNATIONAL JOURNAL OF EDUCATION AND INFORMATION TECHNOLOGIES*, *10*, 229–242.

Alslaity, A., & Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. *Behaviour & Information Technology*, *43*(1), 139–164. https://doi.org/10.1080/0144929X.2022.2156387

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1), 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Brüggemann, P., Lorenz, J.-T., Catlin, T., Chinczewski, J., & Prymaka, S. (2018). Claims in the digital age: How insurers can get started. McKinsey Insights, 1. Retrieved from https://www.mckinsey.com/industries/financial-services/our-insights/claims-in-the-digital-age

Content-based Filtering | Machine Learning. (2022). Google for Developers. Retrieved January 28, 2024, from https://developers.google.com/machine-learning/recommendation/content-based/basics

DesignYuan, S. (2023). Design and Visualization of Python Web Scraping Based on Third-Party Libraries and Selenium Tools. Academic Journal of Computing & Information Science. Vol. 6, Issue 9: 25-31. https://doi.org/10.25236/AJCIS.2023.060904

Dutta, A. (2020). Clustering Techniques and Their Applications: A Review. *American Journal of Advanced Computing*, *1*, 1–6. https://doi.org/10.15864/ajac.1404

Ghasemaghaei, M., & Eslami, S. P. (2016). *Consumers' attitude toward insurance companies: A sentiment analysis of online consumer reviews.*In Proceedings of the Twenty-second Americas Conference on Information Systems. San Diego, CA.

Han, S., & Anderson, C. K. (2021). Web Scraping for Hospitality Research: Overview, Opportunities, and Implications. *Cornell Hospitality Quarterly*, *62*(1), 89–104. https://doi.org/10.1177/1938965520973587

Mitchell, R. (2018). *Web scraping with Python* (2nd ed.). O'Reilly Media., from https://www.academia.edu/41461428/Ryan_Mitchell_Web_Scraping_with_Python_COLLECTING_MORE_DATA_FROM_THE_MODERN_WEB

Marchello, Y., Setiawan, A. W., & Gaol, F. L. (2019). Implementation of K-means clustering for classification of total transaction and seasonal correlation on online retail shop. Journal of System and Management Sciences, 13(6). Retrieved from https://www.aasmr.org/jsms/Archives/Vol-13/Vol-13-6/

Kannengiesser, U., & Gero, J. S. (2023). MODELLING THE DESIGN OF MODELS: AN EXAMPLE USING CRISP-DM. *Proceedings of the Design Society*, *3*, 2705–2714. https://doi.org/10.1017/pds.2023.271

Singhal, A. (2023). K-means clustering algorithm | Examples | Gate Vidyalay. Retrieved July 13, 2024, from https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

Mahesh, B. (2019). *Machine Learning Algorithms -A Review*. https://doi.org/10.21275/ART20203995

Öztürk, F. E., & Demirel, N. (2023). Comparison of the methods to determine optimal number of cluster. *Veri Bilimi*, *6*(1), Article 1.https://dergipark.org.tr/en/pub/veri/issue/78749/1260528

Powell, L., Mccullough, K., Maroney, P., & Cole, C. (2013). Automobile Insurance Vehicle Repair Practices: Politics, Economics, and Consumer Interests. *Risk Management and Insurance Review*. https://doi.org/10.1111/rmir.12032

Rahimi, I., Gandomi, A. H., Chen, F., & Mezura-Montes, E. (2023). A Review on Constraint Handling Techniques for Population-based Algorithms: From single-objective to multi-objective optimization. *Archives of Computational Methods in Engineering*, *30*(3), 2181–2209. https://doi.org/10.1007/s11831-022-09859-9

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58. https://doi.org/10.1145/245108.245121

Richardson, L. (2024). *beautifulsoup4: Screen-scraping library* (4.12.3) [Python]. Retrieved June 20, 2024, from https://www.crummy.com/software/BeautifulSoup/bs4/

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, *9*(1), 59. https://doi.org/10.1186/s40537-022-00592-5

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Computer Science*, *2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Schrotenboer, D. W. (2019). *The Impact of Artificial Intelligence along the Customer Journey: A Systematic Literature Review* [Info:eu-repo/semantics/bachelorThesis]. University of Twente. https://essay.utwente.nl/78520/

Singh, P., & Singh, V. (2024). The power of AI: Enhancing customer loyalty through satisfaction and efficiency. *Cogent Business & Management*, *11*. https://doi.org/10.1080/23311975.2024.2326107

Sirisuriya, S. D. S. (2023). Importance of Web Scraping as a Data Source for Machine Learning Algorithms—Review. *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, 134–139. https://doi.org/10.1109/ICIIS58898.2023.10253502

Vanneschi, L., & Silva, S. (2023). Introduction. In L. Vanneschi & S. Silva (Eds.), *Lectures on Intelligent Systems* (pp. 1–9). Springer International Publishing. https://doi.org/10.1007/978-3-031-17922-8_1

Zhang, Z., Patra, B. G., Yaseen, A., Zhu, J., Sabharwal, R., Roberts, K., Cao, T., & Wu, H. (2023). Scholarly recommendation systems: A literature survey. *Knowledge and Information Systems*, *65*(11), 4433–4478. https://doi.org/10.1007/s10115-023-01901-x

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa