Review for the bachelor’s thesis

## Feedback on Thesis

**Research project**

Completed by a Student in Year 3, group 212 of the Bachelor’s Programme “Data Science and Business Analytics” at the HSE University Faculty of Computer Science

Lesvhev Vladislav Albertovich,

Full Student name

on the topic: Automating the process of selecting customer lists for targeted communication using mathematical modeling: Titova Natalyia Nikolaevna

A work on 49 pages, containing 4 chapters including literature, is represented for review, the list of used sources contains 12 titles.

|  |  |  |  |
| --- | --- | --- | --- |
| No. | **Evaluation Criteria** (only assess those applicable to student’s work)  There is no requirement of research novelty for the BSc student work | **Score (on a 10- point scale)** | **Comments**  *(Here is an approximate list of the competencies developed by the student while completing the thesis. You can refer to them when describing the degree to which the evaluation criteria have been met and place your own comments in the column)* |
| 1 | Research novelty/ Applicability of project results |  | *The practical significance of this work lies in the use of real and current data, enabling the developed recommendation system to be applied in e-commerce for suggesting relevant products to customers. The student explored and implemented modern methods to tackle this task, analyzed challenges encountered during the recommendation system's implementation, and provided conclusions and justifications for which system proved to be the most effective. Given the relevance of this topic, the solutions obtained have numerous potential applications in real practical scenarios.* |
| 2 | Relevance of the work |  | *Some of the critical ways to use customer data effectively in coming up with efficient marketing strategies may help gain a competitive advantage. This further underlines the relevance of the project, since the automations focus on highly labor-intensive processes that are conventionally done through manual approaches. Since the businesses have increasingly developed the need to bring real-time decisions into play along with data analytics, this piece of work meets an important challenge by enhancing operational efficiency and optimizing marketing efforts through automation. This is particularly significant with regard to e-commerce, which is driven by a person's contact and, therefore, generates many records.* |
| 3 | Comprehensive comparative review of known results |  | *The project builds upon well-established techniques such as Exploratory Data Analysis (EDA), pivot tables, clustering algorithms, and MBA. These are recognized tools in the field of data science and customer analytics. However, this project offers a deeper implementation by automating customer segmentation and communication, which often requires substantial manual effort. Compared to traditional marketing analysis, this automated approach offers a more scalable, efficient, and data-driven framework. While similar works focus on either segmentation or predictive modeling, this project combines both aspects into a cohesive solution that optimizes communication strategies.* |
| 4 | Complexity and volume of the completed work |  | *The project demonstrates high complexity through the use of multiple advanced techniques, such as machine learning, clustering, and segmentation. It involves substantial data manipulation and preparation, evident from the creation of the Analytical Base Table (ABT) and various normalization techniques. The project's volume is also considerable, involving a large dataset (over 730,000 rows) with multiple stages, including data cleaning, variable selection, and model validation through techniques like silhouette scores and the Davies-Bouldin Index.* |
| 5 | Quality of the composed text. Clear and structured presentation of ideas. |  | *The text is generally well-structured, with clear objectives, implementation details, and an in-depth explanation of technical processes. The use of diagrams and figures, such as correlation matrices and model validation graphs, strengthens the clarity of the work. However, there are some sections where explanations could be simplified for non-expert readers. For example, certain technical descriptions, particularly around model validation metrics, could be presented in a more digestible manner. Also there was not enough brief information about the project on the page Github, descriptions of each file and more detailed comments.* |
| **FINAL SCORE** | |  |  |

**One benefit** of this work is its well-organized structure and presentational sequence. The author demonstrates that he possesses the methods necessary to carry out the whole analytical cycle of resolving the applied problem of automating the process of selecting customer lists for a targeted communication.

**One drawback** of the report is that the level of **plagiarism** detected is 33%, which exceeds the acceptable limit of 20%. However, this situation occurred because a significant portion of the text related to dataset descriptions and other relevant data was flagged as plagiarism. These sections, while technically marked as borrowed, represent material that is essential for interaction and analysis within the course project itself. Consequently, the higher percentage does not accurately reflect improper use of external sources but rather the nature of the task.

**Strengths of the work:**

**Innovative and Comprehensive Approach:**

The project integrates advanced machine learning techniques such as clustering and Market Basket Analysis, providing a thorough, data-driven solution to optimize customer segmentation and communication strategies.

**Real-World Applicability and Scalability:**

The system's automation of customer list selection has direct applications in industries like e-commerce and retail, offering scalable solutions that enhance marketing efficiency and profitability.

**Strong Theoretical Foundation and Methodological Rigor:**

The project is built on well-established mathematical models and data science techniques, with a robust methodology that includes detailed data preparation, model validation, and insightful business-focused outcomes.

**Weaknesses of the work:**

**High Technical Complexity:**

The project’s extensive use of advanced algorithms and technical jargon makes it challenging for non-expert audiences to follow, which may limit its accessibility for business stakeholders who lack a data science background.

**Limited Practical Implementation and Scalability Discussion:**

While the methodology is strong, the project lacks concrete details on how the system would be implemented in real-world business environments and scaled across different industries or larger datasets. This reduces the clarity on how the models would perform in dynamic business scenarios.

**Assumption of Perfect Data Quality:**

The project assumes the availability of high-quality customer data, but in reality, businesses often face challenges with incomplete or noisy data. A more in-depth discussion on how to handle data quality issues would improve the project's practical relevance.

**In Literature overview there is no texts comparison.**

**Reviewer:**

Position, academic degree, department/place of work Full name Signature

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Contact Email