WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Name

Student ID # 011842196

October 30, 2024

**A. Project Overview**

A shoe company, AeroExpress Shoes, has developed and recently put on the market a new athletic shoe called the Air Nitro. The Air Nitro comes in both a men’s style and a women’s style, and the primary difference between the men’s and women’s styles is foot sizing rather than features. AeroExpress Shoes’ marketing department would benefit from a machine learning model to help them determine who of their previous customers would be most likely to purchase the Air Nitro, thus helping AeroExpress Shoes target their marketing efforts to this population.

**A.1. Organizational Need**

AeroExpress Shoes, in the past, has struggled to use its past customer purchase data successfully in marketing new shoe offerings. The machine learning system will take in data regarding AeroExpress Shoes’ past customers’ purchase history and classify customers as “likely” and “unlikely” purchasers of the Air Nitro. AeroExpress Shoes will then deploy additional marketing campaigns to the “likely” purchasers in order to influence them to purchase this shoe.

**A.2. Context and Background**

AeroExpress Shoes was founded in 2012 as an online-only shoe company specializing in men’s and women’s shoes for formal occasions. The company discovered that the customer base for this niche market was not wide enough to meet desired sales figures, and in 2018, the company rebranded as an athletic and streetwear shoe company, greatly expanding its shoe offerings to include specialty performance shoes for all sports as well as aesthetic streetwear athletic shoes. Since 2018, the company has surpassed its target sales metrics, but research has suggested that with targeted advertising, AeroExpress could raise sales significantly. This ML model will provide a list of previous customers for targeted advertising for the latest Air Nitro shoe offering, and if the targeted marketing campaign is successful, the company plans to pursue additional similar targeted campaigns for future shoe offerings.

**A.3. Outside Works Review**

Outside Work 1:

Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J.,

Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk

of bias in studies on prediction models developed using supervised machine

learning techniques: systematic review. *BMJ*, *375*, n2281.

https://doi.org/10.1136/bmj.n2281

In this article, Navarro et al. review 152 supervised machine learning multivariate prediction models across various medical specialties; the goal of these ML models is to predict various physical health ailments in patients. The authors’ conclusion was that most of these reviewed studies are not methodologically sound for a variety of reasons (for example, the study did not include enough participants, or the study had too much missing data). The authors assert that these flaws in study methodologies create an inherent risk of bias that undermines the validity of the ML model in practice (Navarro et al., 2021).

The authors of this article helpfully underline what is needed to make an effective ML-based study: complete data (i.e., a minimum of missing fields in the data used to train the model), comprehensive data (i.e., involving an appropriate number of participants), and minimal overfitting (i.e., where the ML model has been unintentionally tailored to the training data in such a way that it fails to perform on new data) (Navarro et al., 2021). The authors discuss this in the context of ML studies in the medical realm, but these guidelines inform every usage of machine learning, regardless of context. These principles are equally important and effective in a marketing context (such as the current trial by AeroExpress Shoes) as in a medical context.

The authors’ goal in writing this article is to call out common errors made by medical researchers in employing ML models in order to assist future researchers in identifying ML setups that will result in unintentional bias. In that goal, the authors are successful; they systematically describe the typical parameters of the ML models they review (for example, which algorithms and which statistical validation methods are most often used) and offer advice for improving studies for the future (for example, “Future methodological research could focus on determining the appropriate sample sizes for each supervised learning technique” Navarro et al., 2021). The authors thus effectively contribute to the ongoing conversation of how to create the most effective ML models and how to employ these in a specific context to produce the highest-quality research.

Outside Work 2:

Petković, M., Džeroski, S., & Kocev, D. (2023). Feature ranking for semi-supervised

learning. *Machine Learning*, *112*(11), 4379–4408.

https://doi.org/10.1007/s10994-022-06181-0

The authors’ goal in this article is to provide a discussion of and offer guidance for the use of feature ranking in machine learning. Features are the data points that factor into the ML model’s predictions; for example, if the ML model is built for a medical context to predict the advancement of a disease in an individual, the features may include the individual’s age, weight, height, gender, and medical history. Feature ranking involves making choices as to which features are most important for the ML model’s success. Often, ML models make more accurate predictions once irrelevant or confusing features are removed from the dataset. The authors propose two approaches for feature ranking to optimize the functioning of a supervised ML model (Petković et al., 2023).

The authors point out that this type of guidance is necessary given the current world of big data; there are ever more data points available for researchers to work with, and this creates the need for wisely reducing the dimensionality (i.e., the number of features) of a dataset used for an ML model. Reducing the number of features is important for increasing the accuracy of the model’s predictions, reducing the computational complexity of the model, and facilitating the explanation of how the model works (Petković et al., 2023).

The authors explain in detail the theoretical basis for feature reduction methods, including filter methods, embedded methods, and wrapper methods, as well as their strengths and weaknesses. The authors also discuss which feature reduction methods are most widely employed given a researcher’s chosen ML algorithm (for example, which feature ranking strategies are most effective for ensemble-based algorithms). In addition, since identifying the effectiveness of an algorithm and its output is so critical to this discussion, the authors also discuss in detail evaluation measures for evaluating the accuracy of a model (Petković et al., 2023).

Ultimately, the authors discuss in detail the benefits of feature ranking, offer specific methods (and proof of their validity) for researchers to use when navigating feature ranking, and explain the evaluation and validation of their proposals. Thus, this article is a worthwhile read for any practitioner engaged in creating supervised machine learning models, as feature reduction is a critical step in the model development process.

Outside Work 3:

De Mauro, A., Sestino, A. & Bacconi, A. (2022). Machine learning and artificial

intelligence use in marketing: a general taxonomy. *Ital. J. Mark.* 2022, 439–457.

https://doi.org/10.1007/s43039-022-00057-w

In this article, De Mauro et al. explore the use of machine learning for marketing, offering a comprehensive discussion of the use cases and applications for which ML would be employed in business marketing efforts. The authors offer a systematic review of machine learning applications in business practice and create a guide for the practical implications that marketers should know about (De Mauro et al., 2022). This article is a helpful comprehensive review of the applications and the benefits of machine learning for marketing purposes, and one all business marketers would benefit from reading.

De Mauro et al. seek in this article to create a guide that will be helpful for business marketing practitioners to use, and they contrast their efforts here with previous research work that has addressed these concepts theoretically. They refer to their guidance as “application recipes” for marketing practitioners to use and apply to their own work. They divide their ML use cases into business-facing and consumer-facing use cases; the business-facing applications relate to improving business decision-making and improving the business’s financial outlook, and the consumer-facing use cases relate to improving customer communication and recommendations as well as improving the consumer’s shopping experience (De Mauro et al., 2022).

The guidance of De Mauro et al. in this article is well-researched and highly applicable to business marketers’ needs, and the authors’ approach is systematic, comprehensive, and direct enough to be beneficial to all ML practitioners who plan to use ML for business marketing needs. This article is a great benefit to the discussion of the utility of machine learning for business purposes.

**A.3.a. Description of Outside Works’ Benefits**

Outside Work 1:

Navarro et al. provide helpful information regarding how to avoid bias in creating a supervised ML model and identify where many researchers unintentionally admit bias in their projects (for example, allowing missing data and overfitting their ML model on too-small sample size) (Navarro et al., 2021). This article outlines common mistakes ML model creators often make. Since the AeroExpress project is also a supervised ML project and in many ways very similar to the studies that Navarro et al. review, their recommendations will be used in the AeroExpress model to tailor the dataset and avoid overfitting to ensure relevant and accurate predictions.

The authors also discuss in detail the specific statistical validation methods that their reviewed studies use and how to employ statistical validation to detect bias. This will also be important information for the AeroExpress project manager to use in judging the accuracy and relevance of the AeroExpress ML model after it is built.

Outside Work 2:

This article is especially relevant to the development of AeroExpress’s ML model because the AeroExpress model will be a supervised categorical ML model (i.e., a model with a training set that includes the outcome, where the outcome is predicting “Yes” or “No” as opposed to a continuous value), and this article is a survey of the issues concerning the most widely used supervised categorical ML models.

Since feature reduction will likely be a part of the AeroExpress project, in order to most accurately understand which pieces of the dataset have the greatest impact on the ML model’s predictive output, this article’s guidance will be very helpful in identifying the importance and the most effective process for reducing the number of features in the provided AeroExpress dataset. The authors discuss supervised and semi-supervised learning, which will be highly relevant to AeroExpress’s needs, given that the ML model for this project is a supervised classification model (Petković et al., 2023).

The AeroExpress project manager will find this article instructive both in identifying the best strategies for feature reduction and identifying evaluation methods that will test whether a given feature reduction was effective. This will be necessary to create the most effective predictive model using AeroExpress’s dataset.

Outside Work 3:

This article is especially helpful for AeroExpress’s machine learning project as it speaks to the utility of ML in marketing specifically and the marketing uses for which ML is often employed. De Mauro et al. systematically review ML literature to discuss the use cases of machine learning for marketing purposes, thus providing a guide of best practices for AeroExpress’s ML project manager to follow. The authors condense the marketing applications into four categories: “shopper fundamentals, consumption experience, decision marking, and financial impact” (De Mauro et al., 2022), and all of these categories speak to considerations that will be important for improving AeroExpress’s targeted marketing efforts.

Ultimately, this article provides both the research and the use-case applications to help guide AeroExpress’s project in meaningful ways. The article lays out the specific consumer-oriented and business-oriented applications that will inform this project, and the authors’ discussion of the applications for consumer communication and the shopping experience are especially pertinent. This article will be especially helpful for AeroExpress’s ML project manager in thinking through which ML models and capabilities will be especially helpful in targeting their marketing efforts for the Air Nitro.

**A.4. Solution Summary**

The machine learning model will be a supervised classification model using AeroExpress Shoes’ past customer purchase data to identify “likely” future purchasers of the Air Nitro shoe. The ML model will take in a sampled, cleaned dataset and apply a Support Vector Machine algorithm to divide customers between likely and unlikely future purchasers. The model will be trained on data from customers who have already purchased the shoe to help identify customers who will be future purchasers.

**A.5. Machine Learning Benefits**

This machine learning system will benefit AeroExpress shoes by identifying customers for targeted marketing purposes. In addition, the supervised classification model employed will take in data already easily accessible (since it will come from AeroExpress’s own datasets) and will be able to be run in minimal time on a company computer (as opposed to needing supercomputing capabilities). This model will also be able to be easily adapted for future uses – whether continuing to refine Air Nitro targeted marketing with updated data or using entirely new data to build a targeted marketing program for future shoe offerings.

**B. Machine Learning Project Design**

The tasks and project members for the project are as follows:

Tasks:

* Identify an appropriate classification algorithm
* Clean dataset
* Load the dataset (or a sample) into the classification algorithm model
* Split the dataset into testing and training pieces
* Create a working model for classification
* Refine the working model for greater accuracy (including potentially testing other classification algorithms to compare accuracy)

Project Members:

* Data Analyst/Engineer (Role: cleaning and visualizing dataset for use)
* Marketing Specialist (Role: ensuring data used and ML outputs are reasonable and useful from a marketing perspective; ensuring ML outputs can be operationalized for future translation into a marketing campaign)
* Project Manager/Machine Learning Engineer (Role: managing the project and its timeline; building, training, and testing the ML model)

**B.1. Scope**

In-scope vs. Out-of-scope parameters

In-scope:

* Cleaning data to serve needed project functions
* Creating a list of past customers who will be likely purchasers of Air Nitro to be used in targeted marketing

Out-of-scope:

* Identifying the percentage likelihood of a given customer purchasing the product
* Producing similar analysis or targeted marketing for other products
* Using data outside AeroExpress Shoes’ existing customer purchase dataset
* Creating the marketing campaign itself that will target these likely customers to purchase Air Nitro

**B.2. Goals, Objectives, and Deliverables**

Goals:

* Identify at least 300 likely marketing targets from the list of past customer purchasers
* Incur no additional costs beyond staff time and database/cloud storage time

Objectives

* Create an ML model whose testing set predicts likely purchasers with at least 90% accuracy
* Complete the project and provide a list of likely purchasers within eight weeks of the project's begin date
* Increase overall profit from the shoe by 20%

Deliverables

* A supervised classification machine learning model using a Support Vector Machine algorithm
* Hardware systems used: Lenovo Desktop machines (1 GHz, 64-bit processor)
* Software systems used: Windows 11 OS, Jupyter Notebook 7.2, Python version 3.12.10
* Data: A cleaned dataset derived from AeroExpress’s past customer purchasing data
* A list of past customers who have been identified by the ML model as likely purchasers of the Air Nitro shoe for use in a future targeted marketing campaign

**B.3. Standard Methodology**

This project will employ the SEMMA methodology for building a productive model and iterating through the various pieces in order to assess and improve as the project is in production. SEMMA stands for “sample, explore, modify, model, assess,” and the acronym speaks to the ongoing optimization inherent in the model (Software Testing Help, 2024).

**Sample:** The first phase of SEMMA utilizes the dataset and extracts a sample from the full data that will be representative of the whole. Sampling is employed in this methodology in order to reduce the size of the dataset, thus reducing the time and resources involved in computation (Software Testing Help, 2024). In the AeroExpress ML project, the data analyst/engineer team member will identify an appropriate sample and ensure that it is representative of the full dataset for use in the ML model.

**Explore**: The second phase of SEMMA employs data exploration and cleaning methods to learn more about the data’s trends and ensure that the data is appropriate for the needs of the model (Software Testing Help, 2024). In the AeroExpress ML project, the data analyst/engineer will employ statistical models and tests to check for “dirty data” and ensure there are no outliers that would unduly affect the ML model’s predictions. If outliers are identified, the data analyst/engineer will consider removing them from the dataset.

**Modify:** The third phase of SEMMA involves modifying the data in order to best fit the computational needs of the model (Software Testing Help, 2024). For example, if certain fields are currently represented using continuous data and need to be represented with categorical data in order to be best used for ML needs, that transformation will happen in this phase. In the AeroExpress ML project, the data analyst/engineer and ML project manager will work together to identify what fields might need transformation for best use by the ML model, and the data analyst/engineer will put any transformations needed into effect.

**Model:** The fourth phase of SEMMA involves creating the ML model using the sampled and transformed data (Software Testing Help, 2024). The AeroExpress ML project will involve the ML project manager writing code to implement the ML model and creating a working supervised classification predictive model.

**Assess:** The fifth phase of SEMMA involves assessing the “usefulness and reliability of the constructive model,” including employing real data (Software Testing Help, 2024). The AeroExpress ML project will involve the ML project manager using statistical tests to assess the accuracy of the ML model and also trialing other supervised classification algorithms to test whether other algorithms will produce more accurate results.

**B.4. Projected Timeline**

The expected project timeline is eight weeks. The following tasks are expected to be completed in each week of development:

* Week One:
  + Project team meets (virtually), discusses each member’s responsibilities and projected deliverables
  + Database is identified, and all project team members ensure they have access to the database and that online accounts are working properly
  + Data analyst/engineer requests needed dataset from data marketing team (internal)
  + ML project manager identifies supervised classification models to use and begins the process of mocking up code that will be used later
* Week Two:
  + Data analyst/engineer receives needed dataset from the data marketing team and begins the preprocessing and cleaning steps to ensure data is clean
  + The ML project manager has completed the first round of ML model prototyping
* Week Three
  + Data analyst/engineer finishes the preprocessing and cleaning steps and, together with the ML project manager, samples and explores the dataset, identifying trends and outliers with data visualizations and statistical tests
* Week Four
  + The dataset has now been fully cleaned and explored and is now ready for employment in an ML model
  + The ML project manager utilizes the dataset to train the ML model
  + The ML project manager has a working prototype by the end of this week
* Week Five
  + The ML project team convenes to view a demonstration of the working ML prototype
  + The ML project manager identifies and shares a plan for improving the ML model (potentially using updated data and/or other ML algorithms to improve accuracy)
  + The ML project manager pursues the work laid out to improve the ML model
* Week Six
  + The ML project team convenes to view a demonstration of the updated model with updated statistical tests
  + The ML project manager discusses any additional available options for improving the model’s accuracy
  + The ML model meets the accuracy goals by this stage
* Week Seven
  + The final ML model has been tested at this stage
  + The data analyst/engineer creates a presentation and data visualization deliverables to present the ML findings to stakeholders
  + The ML project team submits a list of names to the marketing team of likely marketing targets for the Air Nitro
* Week Eight
  + Final project sign-off and debriefing
  + Identify steps for improvement for future ML models and ML marketing projects

**B.5. Resources and Costs**

|  |  |
| --- | --- |
| Hardware Resources | 4 Lenovo Desktop machines (1 GHz, 64-bit processor) - $ 0 additional cost |
| Software Resources | Windows 11 OS, Jupyter Notebook 7.2,  Python version 3.12.10 - $0 additional  cost |
| Database (Cloud) Hosting Costs | $313.90 per month x 2 = $627.80 |
| Staff Costs (@ 50%):\*  Data Analyst/Engineer  Machine Learning Engineer  Marketing Specialist | Data Analyst/Engineer: ($80,000/year, for eight weeks, at 50%) $6,153.85  ML Engineer: ($120,000/year, for eight weeks, at 50%) $9,230.77  Marketing Specialist: ($70,000/year, for eight weeks, at 50%) $5,384.62 |
| Staff time of IT marketing data team in procuring internal dataset to transmit to ML team | $490 |
| **Total Cost:** | **$21887.04** |

\* This is not the only project staff will be working on during this time. This project is expected to take 50% of staff time during the eight weeks of production time.

**B.6. Evaluation Criteria**

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Accuracy | Predict likely purchasers with at least 90% accuracy. |
| Timeline | Complete the project and produce a final deliverable (list of likely purchasers) within eight weeks of the project's begin date. |
| Additional profit | Increase overall profit from Air Nitro over similar products by 20%. |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

The hypothesis of this project is that the ML project team will be able to predict future purchasers of the AeroExpress Air Nitro shoe using past purchasing data from other customers who have also purchased the Air Nitro shoe. The hypothesis involves the presumption that the past purchase history of customers who purchased the Air Nitro will be similar to the past purchase history of customers who have not yet purchased the Air Nitro but are likely to do so.

**C.2. Selected Algorithm**

The ML model will use a supervised classification algorithm to predict which customers will be likely to purchase the Air Nitro shoe based on their past purchasing history. In this project, the team will use a Support Vector Machine algorithm (SVM) to separate potential customers into likely and unlikely candidates for future purchases. An SVM algorithm is one that separates items into groups using a hyperplane to identify the maximal distance between the two (Avjyan, 2018).

**C.2.a. Algorithm Justification**

The support vector machine algorithm (SVM) was chosen because of its speed and simplicity: it is able to distinguish in a straightforward way between two groups, as this project aims to do.

**C.2.a.i. Algorithm Advantage**

One major advantage of using the SVM algorithm is that it is less prone to overfitting than other machine learning models like decision trees.

**C.2.a.ii. Algorithm Limitation**

One disadvantage, however, is that SVM cannot navigate missing values. In this ML project, the data analyst/engineer will need to ensure that the dataset is complete in all fields (GeeksforGeeks, 2023).

**C.3. Tools and Environment**

The ML project will be implemented using a Windows 11 operating system running Python code in Jupyter Notebooks. SVM algorithms require the use of Numpy, the Python library that facilitates numeric transformations. The project will also require the Scikit-learn library SVC (Support Vector Classification) (Avjyan, 2018).

**C.4. Performance Measurement**

The performance will be measured in the following ways:

* The dataset containing will be split into two parts: 65% for training and 35% for testing.
* The 65% will be used to train the model to understand how to classify purchasers (as likely or unlikely to buy the Air Nitro).
* The model will then run on the testing group. A series of statistical tests can be performed on the testing group to determine how accurate the model is. These tests include:
  + Confusion matrix, which shows the true positives, true negatives, false positives, and false negatives provided by the model
  + Accuracy, precision, and recall (sensitivity), which are metrics derived from the values provided by the confusion matrix
  + F1-score, which combines precision and recall to provide another way of conceptualizing the percentage of true positive results

The metric being calculated (and which is being measured) is the accuracy of the machine learning model in predicting which customers will purchase the Air Nitro. The goal for the ML model is to achieve 90% accuracy (representing the ratio of correct predictions to incorrect predictions, as identified by the confusion matrix and derivative metrics).

.

**D. Description of Data Sets**

**D.1. Data Source**

The dataset used for the machine learning model will come from AeroExpress’s own past customer data. It will include unique customer IDs and the customer’s history of purchases from AeroExpress. Since AeroExpress also employs cookies on its website to track customers’ movements online, the dataset will also include information regarding competitors’ websites the customer has also visited.

**D.2. Data Collection Method**

The dataset will come from the data AeroExpress already collects and has available on its servers. Procuring the data for this project will be simply a matter of making an internal request to the IT team to provide the needed data.

**D.2.a.i. Data Collection Method Advantage**

One advantage of this data collection method is that it is readily available and free to use for internal AeroExpress purposes like this machine learning project.

**D.2.a.ii. Data Collection Method Limitation**

One disadvantage is that because the dataset includes only data collected by AeroExpress, the data is relatively incomplete. The dataset will not include information regarding purchases on other websites or details regarding other aspects of the customers’ consumer habits.

**D.3. Quality and Completeness of Data**

The dataset used for the machine learning project will include a unique customer ID, AeroExpress products previously purchased by the customer, and other websites the customer has visited. Some of the customers will also have purchased the Air Nitro shoe; there will be a boolean field indicating whether the customer has also already purchased the Air Nitro shoe. The unique customer ID is automatically assigned upon the customer’s first website purchase and is an integer. The past customer purchases are represented in the dataset by a list of the unique IDs of the products purchased. The other websites visited by the customer are represented by a list of the website’s home URLs.

The dataset used for the machine learning project will be an extract of the total data available in AeroExpress’s database; it will not include customers’ personal details like name, address, etc., as these are not relevant to the project. Customers will be identified in the project only using their unique auto-assigned ID number. Therefore, there will not be concerns about the handling of personally sensitive information or bias on the part of the project manager.

There will be several data cleaning considerations that will need to be taken into account before the data will be ready for the machine learning model. First, it is possible that the same customer will have made past purchases using different accounts using different email addresses. It will be important to identify from the customer information fact tables (not used by the machine learning project but available in AeroExpress’s data collection) whether there are multiple accounts that can be tied to the same person. For the machine learning project, these multiple accounts will need to be coalesced into the same account in order to have a holistic view of that customer’s past purchasing and web movement habits.

Second, the data will need to be validated to ensure that URLs used for the customers’ visited pages all use the same format for a given visited webpage. For example, the team will need to ensure that https://www.zappos.com, http://www.zappos.com, http://zappos.com, and zappos.com are all encoded in the data in the same way. This will be necessary in order for the ML model to recognize that these URLs refer to the same site.

The start dates and end dates for the information contained in the dataset are an important consideration. AeroExpress releases approximately five new styles of athletic shoes per year, and the company transitioned significantly from a focus on formal dress wear to a focus on athletic and streetwear in 2018. Therefore, the cutoff “start” date for tracking customers’ past purchases will be January 1, 2018. In addition, the ML model will benefit from training on a sample of the customers who have already purchased the Air Nitro. It will be helpful to refresh the data on a monthly basis to add to the dataset customers’ updated purchase information.

**D.4. Precautions for Sensitive Data**

The dataset used for the ML model itself will not contain sensitive data since customers will be identified only by an auto-assigned integer Customer ID. However, the Data Analyst/Engineer on the project who will take care of data cleaning and validating will also need to use other internal AeroExpress customer datasets that contain sensitive information. The Data Analyst/Engineer will not need additional data access since this will be a regular part of his or her job responsibilities. The Data Analyst/Engineer will provide only the cleaned, validated, non-sensitive dataset for the ML model to the Project Manager.

It is AeroExpress’s policy regarding data not to transmit anything potentially sensitive over email. Any transmission of data must take place using the company’s internal storage server. This will reduce the risk of interception by outside actors. Even though the transmitted dataset does not contain sensitive information, it would still be valuable to AeroExpress’s competitors. In a similar vein, AeroExpress requires its employees to sign non-compete agreements, ensuring they will not work for other shoe companies during their tenure at AeroExpress and will not store any information on personal computers.

**References**

Avjyan, H. (2018, November 15). *Overview of supervised learning model SVM (Support*

*Vector Machines)*. Medium. https://medium.com/@hakobavjyan/overview-of-

supervised-learning-model-svm-support-vector-machines-20b683a4eaf

De Mauro, A., Sestino, A. & Bacconi, A. (2022). Machine learning and artificial

intelligence use in marketing: a general taxonomy. *Ital. J. Mark.* 2022, 439–457.

https://doi.org/10.1007/s43039-022-00057-w

GeeksforGeeks. (2023, May 7). *Support Vector Machine in machine learning*.

https://www.geeksforgeeks.org/support-vector-machine-in-

machine-learning/#

Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J.,

Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk

of bias in studies on prediction models developed using supervised machine

learning techniques: systematic review. *BMJ*, *375*, n2281.

https://doi.org/10.1136/bmj.n2281

Petković, M., Džeroski, S., & Kocev, D. (2023). Feature ranking for semi-supervised

learning. *Machine Learning*, *112*(11), 4379–4408.

https://doi.org/10.1007/s10994-022-06181-0

Software Testing Help. (2024, May 10). *Data Mining Process: Models, Process Steps &*

*Challenges Involved*. https://www.softwaretestinghelp.com/data-mining-

process/#Data\_Mining\_Models