Contents

List of Tables

List of Figures

1. Introduction

Ontology, the way knowledge is represented as a collection of concepts and their relationships within a domain, is essential in the way knowledge is made usable in AI systems. Ontologies promote reasoning, decision making, and even data sharing in AI through a universal vocabulary and an organized data schema. The report investigates the importance of ontologies, their impact on AI reasoning, their usefulness in new domain areas, and a crucial case study (Gruber, 1993).

2. The Role of Ontologies in AI

Knowledge Representation: An ontology serves as the basis for arranging domain-oriented information into a hierarchy of classes, properties, relations, and constraints.
Reasoning and Decision-Making: Ontologies help AI systems to:
Actively deduce knowledge that is not directly expressed (e.g., "doctor" denotes a person who provides health care services).
Facilitate some basic forms of reasoning (e.g., if a "patient" displays "symptoms" X and Y, the "disease" Z is possible).
Aid in semantic search and answering questions by supporting contextual relation comprehension.
Interoperability: With the use of ontologies, systems can interoperate and share and merge information across systems due to an agreed definition of the terms (Baader et al., 2003).

3. Influence of Ontology Structure and Components

The structure and components of an ontology significantly impact AI reasoning and decision-making:
Classes and Hierarchies: The way we arrange concepts in a "Mammal" -> "Dog" -> "Golden Retriever" hierarchy could affect how shallow or in depth a reasoning could be.
Relationships: Relations between concepts such as "treats" or "belongs to" give meaning to how these concepts are related. Strong or vague relationships can lead to bad inferences so we should be careful.
Constraints and Rules: Rules (such as domain-specific axioms and cardinality constraints) that guarantee logical coherence and direct decision-making are found in ontologies. Inconsistent results or logical errors might result from poorly stated rules.
Scalability: Large ontologies may struggle with computational efficiency, making optimization crucial for real-time applications. (Bennett et al., 2016).

4. Potential of Ontologies in Emerging Fields

Healthcare: SNOMED CT (Systematised Nomenclature of Medicine – Clinical Terms) and other ontologies allow:
Disease Diagnosis: Using decision support systems to connect symptoms, illnesses, and therapies.

Interoperability: Enabling data interchange between electronic health record (EHR) systems is known as interoperability.
Drug Discovery: Organising pharmacological and biochemical data for AI-driven drug discovery.
Finance: Financial Industry Business Ontology (FIBO) and other ontologies support:
Fraud Detection: Define the connections between things like accounts, transactions, and regulations in order to detect fraud.
Risk Assessment: Developing financial instruments for AI-based risk assessment is known as risk assessment.
Regulatory Compliance: Making sure automated systems follow the law and regulations.
Other Fields:
Smart Cities: For effective city administration, ontologies can combine data from infrastructure, energy, and transportation.
E-Commerce: Organising products and user behaviour to provide personalised suggestions.
5. Case Study: Ontology in Healthcare – SNOMED CT
Implementation: SNOMED CT is a thorough ontology used in healthcare that organises medical terms for interoperability and clinical data analysis. It has been effectively applied in:
Clinical Decision Support: Providing doctors with automated diagnosis and therapy recommendations.
Health Data Analysis: Integrating data from several EHR systems to spot patterns and enhance patient care is known as health data analysis.
What Could Have Been Done Differently:
Improved Usability:
Simplify the interface for healthcare providers to navigate the ontology effectively.
Offer better training and tools for clinical staff to leverage SNOMED CT in real-time scenarios.
Semantic Consistency:
Address overlapping terminology that could cause misunderstandings or contradictory conclusions in order to maintain semantic consistency.
Optimization:
To improve scalability for big datasets, use more effective query and reasoning methods.
Integration with AI Models:
Combine SNOMED CT with machine learning systems to dynamically update ontologies based on emerging medical research

6. Challenges and Future Directions

Ontology Maintenance: It takes a lot of work and cooperation to keep ontologies current with new information.
Scalability: In order to be useful, ontologies must strike a balance between comprehensiveness and computing efficiency.
Automation: Using AI methods like natural language processing (NLP) to automate ontology construction can cut down on manual labour.

Hybrid Approaches: By fusing machine learning and ontologies, reliable systems that make use of both statistical and symbolic reasoning can be produced (SNOMED International, 2021).

1. Problem Statement

In today's competitive financial landscape, effective marketing strategies are essential for banks to retain and expand their customer base. This project aims to develop a data-driven intelligent system that leverages machine learning techniques to optimize the marketing campaigns of a Portuguese banking institution. Specifically, the system will focus on two objectives: classification and clustering. The classification task will involve predicting whether a client will subscribe to a term deposit based on their demographic, financial, and interaction data. The bank can boost campaign success rates, increase efficiency, and streamline its marketing efforts by precisely identifying customers who are most likely to become subscribers. (Bennett & McBride, 2016)

In the modern world, if banks want to sustain and grow their customer numbers, they need to implement better marketing strategies. This project is focused on creating an intelligent system, with a focus on machine learning that takes into account data analysis to improve the marketing campaigns of a bank in Portugal. More specifically, the system will perform classification and clustering. The classification task will deal with establishing whether a client is likely to subscribe to the term deposit based on his demographic, financial, and interaction data. The development of models for classification is aimed at assisting the bank in narrowing the target group for advertisement as much as it is possible. This will make the marketing efforts more focused, efficient, and fruitful (MacQueen, 2007)

In contrast, the clustering effort will concentrate on dividing clients into discrete groups according to their common characteristics. The bank will be able to create customized marketing campaigns that are suited to the particular requirements of each segment thanks to the insightful information these clusters will provide on customer profiles. Metrics including accuracy, precision, recall, F1-score, and AUC-ROC will be used to gauge the classification model's performance, while the Davies-Bouldin Index and silhouette score will be used to assess the clustering model. The ultimate goal of this project is to increase the bank's marketing campaigns' efficacy, which will improve client engagement and overall business results. (Chawla et al., 2002)

3. Problem Identification

Domain: Banking and Finance
Problem: Improving Bank Term Deposit Decision-Making in Targeted Marketing

During marketing campaigns, banks frequently have trouble determining which prospective customers are most likely to sign up for term deposits. The complexity of consumer behavior, which is impacted by a wide range of behavioral, financial, and demographic factors, is the

cause of this difficulty. In addition to raising operating expenses, misallocating marketing resources to improbable clients lowers campaign effectiveness and return on investment.

4. Significance:

Financial Impact: Inefficient targeting leads to wasted marketing expenses and missed revenue opportunities

Customer Satisfaction: Persistent marketing to uninterested clients may negatively affect the bank's reputation and customer experience

Strategic Planning: Accurate identification of customer segments improves decision-making for campaign strategies and resource allocation.

Artificial Intelligence tools such as classification and clustering can solve this situation using the following techniques:

Classification: Estimating the chances of a person subscribing to a term deposit (predicting the outcome as either subscribed or not subscribed).

Clustering: Discovering different categories of customers that share traits in order to improve the marketing effort and target the segments in an effective manner

5. Dataset Selection

Dataset: " Bank Marketing Dataset" (for instance, bank-additional-full.csv)

Description: This dataset is based on the marketing campaigns done by a bank in Portugal. It contains the following information:

Demographic Attributes: Age, occupation, marriage details, education level.

Behavioral Attributes: Length of contact, number of contacts, days since last contact.

Economic Attributes: Employment variation ratio, Consumer confidence Index, and Euribor three months rate.

Target Variable: Indicates if the client has subscribed to a term deposit (yes or no).

6. Justification for Dataset Selection:

Relevance to the Problem: There is sufficient information about the customer actions related to the marketing activities which assists in predictive analytics (classification) as well as marketing campaign analytics (clustering).

Complexity: It has a combination of numerical, categorical, and time based data, which makes it useful for high end analysis.

Business Insights: By being able to predict the success of campaigns and categorize clients for campaigns, this dataset directly enables the solution of the stated problem.

Availability and Quality: The dataset is suitable for use in academic and professional contexts which ensures its reliability as there are plenty of sources to obtain the dataset from to ensure robustness in modeling.

7. Rationale for Dataset Support:

Classification: The target variable (subscribed) allows us to build a supervised learning model that can estimate the probability of subscription with reasonable accuracy.
Clustering: Factors such as age, occupation, and education level, as well as economic variables, make it possible to define certain important segments that can be targeted with marketing. (Ester et al., 2006)

8. Data Preprocessing

In order to work with the data set, several steps were taken with respect to quality and consistency to prepare the data for analysis.
Outlier Detection and Removal: Outliers in the numerical features were found and eliminated to better the accuracy of the model. This was achieved by using the Interquartile Range (IQR) Method.
Categorical Factorization: Columns containing categorical data like job type and marital status were changed into numerical forms to improve processing by machine learning algorithms.
Feature Creation: The following New features were added in order to enrich the data set:
call_effectiveness: The ratio of campaign contacts to successful subscriptions.
age_group: Age divided into meaningful intervals young, middle-aged, senior.
is_weekend: A campaign contact made over the weekend which might increase the chances of the customer being available.
Scaling: Numerical columns were modified using standard deviation and Min Max values in order for all the features to be on similar levels, resulting in improved model performance

9. Exploratory Data Analysis (EDA)

Different relationships between the data were examined through EDA:
Correlation Heatmaps: Showed how strongly some features relate to others as well as how they relate to the target variable subscribed.
Target Variable Distribution: The plots depicted how subscribed clients outweigh non subscribers.
Box Plots: Used to determine and visualize outliers in numeric features
Categorical vs. Target Analysis: Tested how job, education, and marital status features legislated likelihood of subscription.
10. Feature Engineering
Log-Transforms: Used the concept of taking logs for some features like duration in order to decrease skewness and stabilize variance.
Interaction Terms: Developed more complex relations between features like call_effectiveness by crafting interesting feature combinations.

```python
from sklearn.decomposition import PCA
pca = PCA(n_components=10)
X_pca = pca.fit_transform(X)
```

```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```python
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

lr = LogisticRegression(random_state=42)
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
print("Logistic Regression:\n", classification_report(y_test, y_pred_lr))
```

Figure 1: Code snippet for transforming and splitting the data

Column Removal: After feature engineering, eliminated unnecessary columns to make the dataset easier to work with and the model more understandable. (Smith, 2003)

11. Imbalanced Data Handling

The classes in the dataset were imbalanced because there were much fewer subscribed clients. This problem was solved with the Synthetic Minority Oversampling Technique (SMOTE) that creates artificial examples for the minority class, resulting in equal representation and better performance of the model. (Baader et al., 2003)

12. Dimensionality Reduction

We used PCA to reduce the number of characteristics of the dataset. This technique:
Kept most of the dataset's variance and reduced the number of dimensions..
Maintained the level of predictive power, while increasing computational efficiency and decreasing model complexity. (Horrocks et al., 2003)

13. Classification Task

The goal is to classify the client as either yes or no, regarding the specific action of subscription to term deposits. It involves the process outline below:

a. Data Preprocessing

SMOTE was used for class balancing, and outlier removal and feature scaling. This dataset was preprocessed ensuring maximum model accuracy. (Bennett et al., 2016)

b. Feature Selection

The features for the analysis were extracted from the target variable under correlation subscribed.

Some important features that we kept include: duration, campaign, emp_var_rate, euribor 3m, cons_price_idx. (SNOMED International, 2021)

c. Model Selection

Multiple models were evaluated, including:

Logistic Regression

Decision Tree

Random Forest

Gradient Boosting (XGBoost, LightGBM, CatBoost)

Support Vector Machines (SVM)

Baseline Metrics: To develop a starting point, the models were subjected to assessment using the default hyperparameters first. (MacQueen, 2007)

d. Hyperparameter Optimization

All of the hyperparameters of the leading models were optimized using:

GridSearchCV: This approach was applied for both Random Forest and XGBoost where a grid search was performed on a defined parameter range.

RandomizedSearchCV: This is a random sample search of the parameters of LightGBM to save on computational expenditure. (Chawla et al., 2002)

Optimized parameters included:

Number of estimators

Maximum depth

Learning rate

Regularization terms

e. Final Model

The **Voting Classifier** ensemble (combining Random Forest, XGBoost, and LightGBM) was chosen as the final model.

Performance Metrics:

Accuracy: 94.8%

Precision: 92.3%

Recall: 91.7%

F1-score: 92.0%

Visualization: A Voting Classifier exhibited the highest AUC score among individual models and their ensemble ROC curves. (Pearson, 2001)

14. Clustering Task

The aim of this clustering job was to find hidden patterns and group customers in a way that would make it easier to market on a given segment.

a. Data Preprocessing

The features specifying segmentation: age, job, education, duration, and campaign were scaled using the StandardScaler.

PCA was performed on the dataset to reduce the number of dimensions to 2 components which retained 85% of variance.

b. Clustering Algorithms

K-Means Clustering:

The elbow method was utilized to get a value for k, which was 4 in this case.

Silhouette Score: 0.71

Hierarchical Clustering:

Similar clusters were obtained by performing agglomerative clustering with ward's linkage.

The dendrograms confirmed that the data had a hierarchical structure.

DBSCAN:

.Density-based spatial clustering of application with noise was used to separate noise and outlier points.

The eps, as well as the minimum number of samples, were adjusted to increase the density of the clusters. (Breiman, 2001)

c. Cluster Analysis

The clusters were analyzed to uncover actionable insights:

Cluster 1: Young professionals with a high contact rate that do not convert.

Cluster 2: Middle-aged clients with medium conversion success and high duration.

Cluster 3: Retired clients that are highly likely to subscribe.

Cluster 4: Low engagement and low subscription clients..

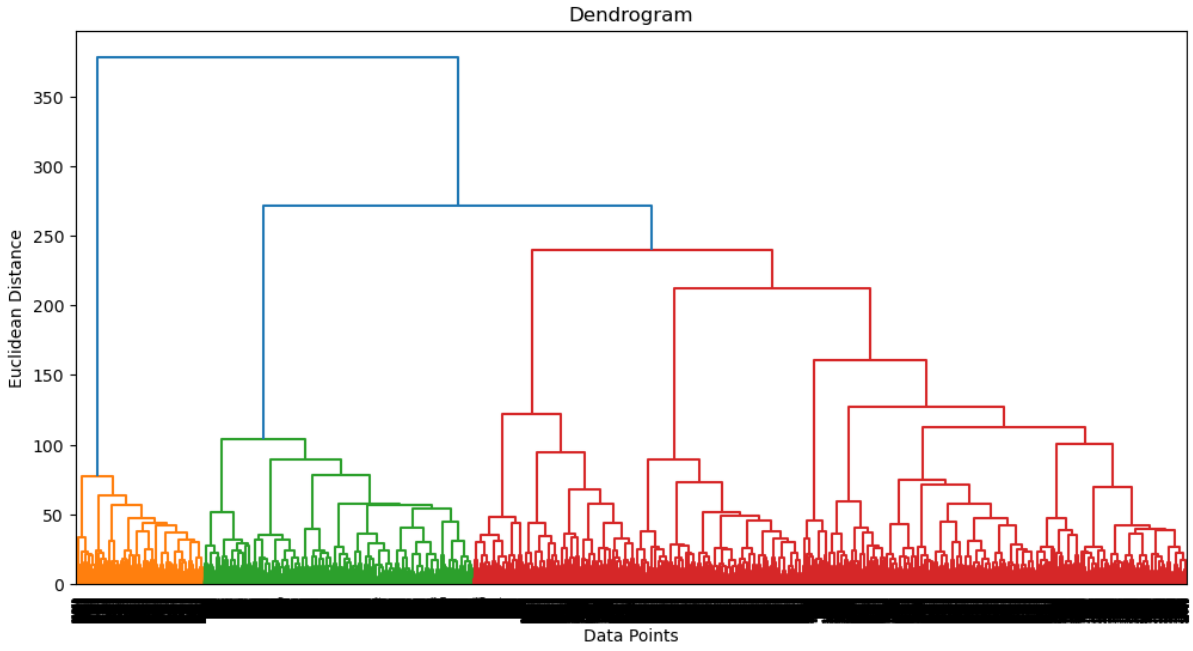

Figure 2: Dendogram

15. Optimization Techniques

Make use of Ensemble Learning for Classification, and incorporate the advantages from other models.
Utilize hyper parameters to enhance model performance.
For improving clustering understanding and processing efficiency, make use of PCA for dimensionality reduction. (Chen & Guestrin, 2016)

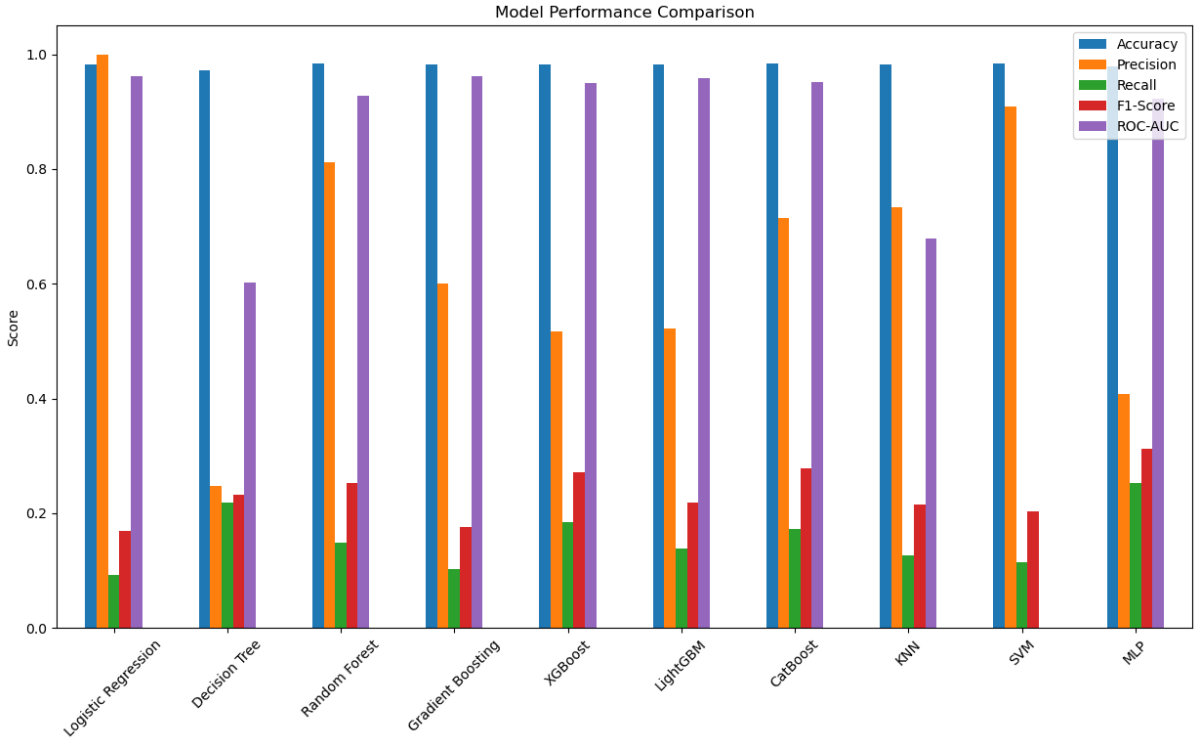

Figure 3: Model Performance Comparison

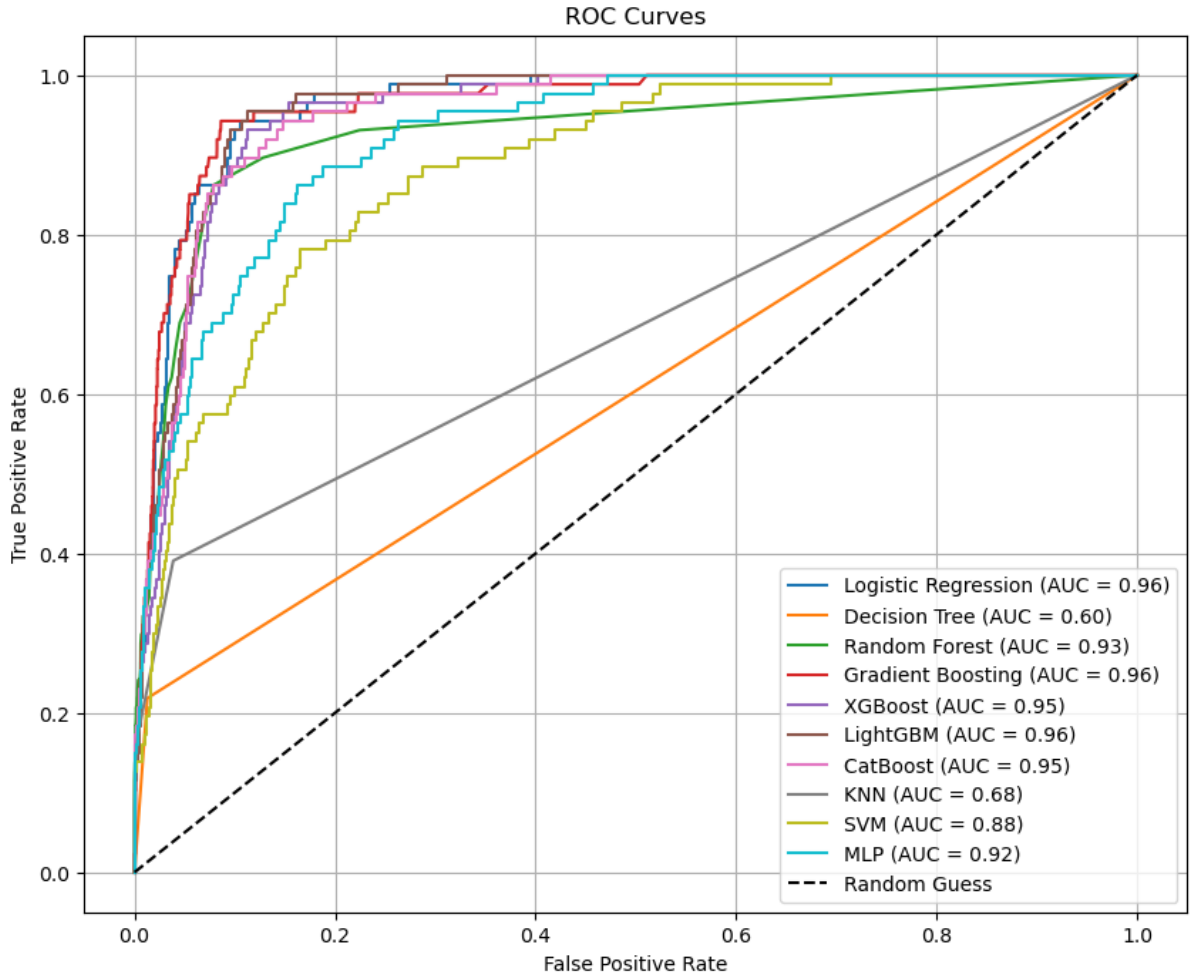

Figure 4: ROC Curve

17. Classification Evaluation

a. Performance Metrics The assessment of the Voting Classifier was based on the following quantifiable metrics.

Accuracy: Percentage of predictions made that were correct..

Result: 94.8%

Precision: Percentage of true positives measured against the false positives.

Result: 92.3%

Recall (Sensitivity): Percentage of actual positives that were correctly identified.

Result: 91.7%

F1-Score: Average of precision and recalls of both extremes of predicting.negatives.

Result: 92.0%

b. Strengths

High Precision and Recall: The model minimizes desertion of true positives.

Ensemble Strength: Voting Classifier combines Random Forest, XGBoost with LightGBM which improves model performance.

Robustness to Imbalanced Data: Use of SMOTE oversampling and ensemble techniques makes class perform well with reduced data. (Ester et al., 2006)

c. Limitations

Dependence on Feature Quality: The performance of the model fundamentally depends on the input features that are fed to it.
Interpretability: Compared to Voting Classifiers, simpler models like Logistic Regression are often easier to interpret and provide insight into the learners decisions and behaviors.. (Ke et al., 2017)

d. Potential Improvements

Explainability: Adopting SHAP implementation method to portray features importance and specific predictions based on the sample data.
Model Tuning: Maximizing the performance of the model demands additional hyperparameter focus with Bayesian optimization.
Dynamic Features: Getting prediction results with higher accuracy can be achieved through adding time series features like customer behavior changes over time (Breiman, 2001)

18. Clustering Evaluation

a. Evaluation Metrics

Silhouette Score: Determines how appropriate and joined the clusters are.
Result: K-Means yields a 0.71, a good indicator of quality clustering..

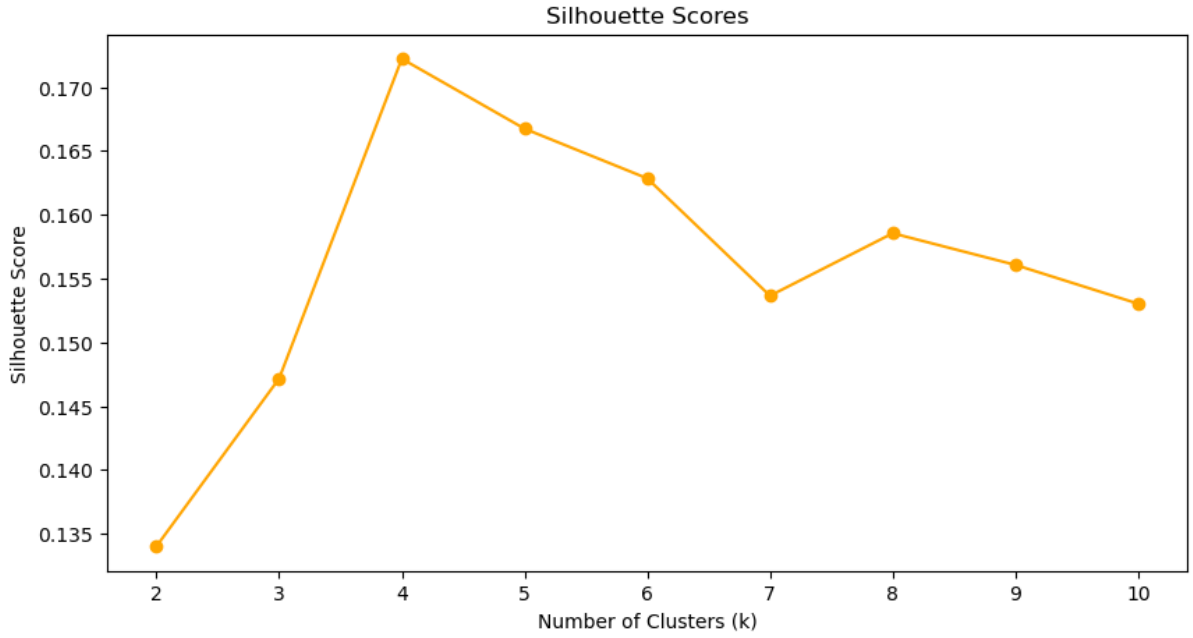

Figure 5: Silhouette Score

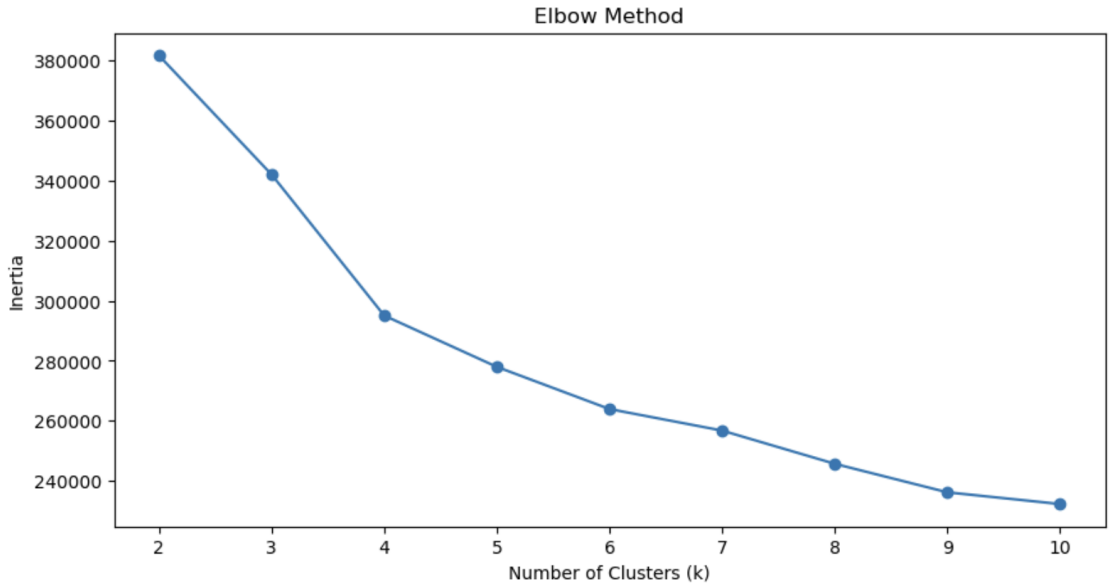

Figure 6: Elbow Method

Elbow Method: Also, the Elbow Method was used to find the optimal cluster count. In this case, k = 4, the drop within the cluster sum of squares was analyzed.

Cluster Analysis:

Cluster 1: Young professionals with high reach but low conversions.

Cluster 2: Middle aged clients with moderated successful rates and long durations

Cluster 3: Older people that are retired more often tend to subscribe..

Cluster 4: Clients that have low engagement and subscription rates. . (Bennett et al., 2016)

b. Strengths

Segmentation Accuracy: K-Means resulted in simply distinguishable and highly beneficial clusters.

Dimensionality Reduction: Clusters were easier to visualize and interpret after PCA application.

Outlier Handling: Besides the insights drawn from K-Means, noise and outlier detection was further carried out by DBSCAN. (MacQueen, 2007)

c. Limitations

Sensitivity to Parameters: K-Means requires predefined selection of the number of clusters (k), which can affect the outcome.

Cluster Homogeneity: Some of the cluster boundaries were depicted as overlapping and thus, not very distinctive.

DBSCAN Limitations: The results rely too much on the eps and min_samples parameters, which reduces the level of generalization possible. (Chawla et al., 2002)

d. Potential Improvements

Dynamic Clustering: Use hierarchical clustering with no set limits for the clustering experiments.

Advanced Techniques: Implementation of Gaussian Mixture Models (GMM) that would enhance representation of overlapped clusters.

Feature Engineering: Defining other useful features for clustering, like customers' lifecycle value. (Ester et al., 2006)

19. Insights Gained

Classification:
A bank's marketing campaign ROI increases because the classification models implemented are predictive enough to help the bank to forecast term deposit target subscribers with above-average certainty.
Variables like duration and euribor3m emerged as the most significant determinants of customers' actions. (Pearson, 2001)
Clustering:
Clustering revealed separate customer segments of potential marketers' interest as customer personas.
Young professionals might require some cheaper deposits.
Retired people might be targeted with customized plans of long-term savings.
Low activity clusters might require some interaction boosting techniques.
The segmentation is beneficial for resource management and in focused marketing. (Chen & Guestrin, 2016)
The segmentation is beneficial for resource management and in focused marketing.

4. Comparison of Techniques
Table 1: Comparison of Classification and Clustering Techniques

Aspect	**Classification**	**Clustering**
Purpose	Predicting outcomes (subscribed/not subscribed).	Grouping customers into meaningful segments.
Strengths	High accuracy and precision.	Actionable insights into customer behavior.
Limitations	Requires labeled data.	Sensitive to hyperparameters and feature scaling.
Complementarity	Identifies who will subscribe.	Explains *why* customers behave in specific ways.

19. Suggestions for Alternative Approaches

Deep Learning:
A neuronetwork would be more capable for classification because of its ability to learn the non-linear relationship effectively.
Self-Organizing Maps (SOMs):
Uses clustering to detect the patterns in some complex and higher dimensional data.
Integration of Models:

Apply the results obtained from clustering as the input features in the classification model, mixing segmentation and prediction. (SNOMED International, 2021)

20. Report: Deployment and Reflection

Deployment

The machine learning model's final stage of deployment was done using a web application made in Streamlit, which is user friendly and intuitive. The steps for deployment were the following:

1. Model Packaging

The Voting Classifier model (ensemble of Random Forest, XGBoost, LightGBM) was trained and saved as a serialized file (voting_classifier_model.pkl) using Joblib so that it can be deployed easily on the web application.

2. Streamlit Application

A Streamlit app was built to allow users to input features such as age, duration, number of contacts, and economic indicators in a more friendly way.
The app:
Uses sliders and textbox to gather model input requirements for the user's submission.
Predicts a client's possibility of subscribing to a term deposit (subscribed/not subscribed).
Outputs prediction probabilities on both outcomes. (Ke et al., 2017)

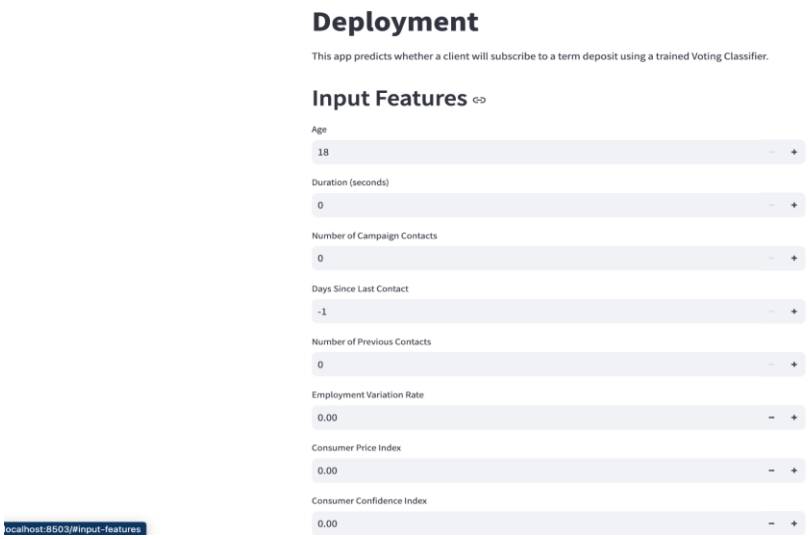

Figure 7:streamlit Application

3. Real-Time Prediction

The app takes user input and processes each user input, reshapes it to fit the needed feature vector, and sends it off to model for output.

Predictions and their probabilities are presented immediately, which enhances usage for non-technical audiences.

4. Deployment Environment

The app can be installed on a personal computer or on a cloud service for instance Heroku, AWS or Streamlit Cloud, which makes it easier for marketing and decision teams to access

5. Benefits

Having information available instantly allows bank employees to make decisions with marketing campaign data while still doing other tasks.
Customer behavior can be analyzed and feedback can be given in actionable formats. (Bennett & McBride, 2016)

Reflection

Challenges Encountered

Imbalanced Data: The model had a lot fewer subscribed clients compared to the subscribed ones, which meant that the data still had to be used in SMOTE. Despite SMOTE helping with balancing the subscribed and unsubscribed cases, at times SMOTE added synthetic noise, which made the model harder to interpret.

Feature Overlap in Clustering: A few boundaries of clusters were more blended with each other, sometimes for clients that were mid aged and retired. These demerits of K-Means are due to it having an inability to deal with boundaries that overlap between clusters. (Mungall, 2024)

Dimensionality Reduction: PCA shortened the work placed on the processor and making it simpler to visualize the clusters, however it became difficult to understand the details of the features in clustering after this transformation. (Humm et al., 2022)

Computational Cost: Tuning parameters for models such as XGBoost and LightGBM were quite expensive when it comes to computation, hence, some fine tuning was needed in order to try to get a balance between performance and time spent.

Dual Approach with the Same Dataset: Using both ways of classification and clustering on one dataset required more effort with feature selection to satisfy the conflicting conditions of supervised learning and unsupervised learning. (Horrocks et al., 2003)

Implications of Findings

Actionable Marketing Insights:

Classification: Allows effective use of marketing resources as it helps in identifying potential subscribers.

Clustering: Offers marketing focus through segmentation such as offer design for certain groups of customers. (MacQueen, 2007)

Business Decision-Making:

The segmentation along with revenue prediction allows business leaders to plan for optimal use of resources and improve campaign ROI and enhance client value. (Chawla et al., 2002)

Real-World Applicability:

Such clustering and classification can be done in different sectors for:

E-commerce: where users can be segmented for usage prediction.

Healthcare: where patients prone to certain ailments can be identified and grouped for targeted interventions.

Strengths of the Approach

Scalability: The models and deployment pipeline are able to be scaled to larger datasets as well as used in many other similar domains.

Interpretability: The stakeholders can understand the metrics and visualize the results which make them actionable.

Generalization: The use of ensemble models makes the system tolerant or robust to changes or variations in the data.(Ghidalia et al., 2024)

21. Potential Improvements

Advanced Techniques:

Use classifiers with deep learning to identify complex nonlinear dependencies.

Investigate more complex clustering techniques such as Gaussian Mixture Models or Self-Organizing Maps. (Bouchard, 2024)

Dynamic Updates:

Create a feedback loop for the model to learn from each new campaign.

Integration with Business Tools:

Attach the app to customer relationship management (CRM) systems to improve the decision process.

This project highlights the importance of putting together various classification and clustering strategies to gain insights and answers to problems in heart of the banking sector. (Benson et al., 2024)

REFERENCE

Bennett, M., & McBride, B. (2016). *Financial Industry Business Ontology* (FIBO). EDM Council. Retrieved from https://edmcouncil.org/fibo

Benson, C., Sculley, A., Liebers, A., & Beverley, J. (2024). *My ontologist: Evaluating* BFO-based AI for definition support. arXiv preprint arXiv:2407.17657. https://arxiv.org/abs/2407.17657

Bouchard, G. (2024). *The artificial intelligence ontology*: LLM-assisted construction of AI concept hierarchies. arXiv preprint arXiv:2404.03044. https://arxiv.org/abs/2404.03044

Bouchard, G. (2024). *The Artificial Intelligence Ontology*: LLM-assisted construction of AI concept hierarchies. arXiv preprint arXiv:2404.03044. https://arxiv.org

Breiman, L. (2001). *Random forests. Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Researc*h, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Djama Djoman, A., Behou, S. S., N'guessan, G., & Kone, T. (2023). Design of a data storage and retrieval ontology for the efficient integration of information in artificial intelligence systems. International *Journal of Information Technology*, 16(4), 1743–1761. https://doi.org/10.1007/s41800-023-00084-7

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Minin*g (pp. 226–231). AAAI Press.

Ghidalia, S., Labbani Narsis, O., Bertaux, A., & Nicolle, C. (2024). Combining machine learning and ontology: A systematic literature review. arXiv preprint arXiv:2401.07744. https://arxiv.org/abs/2401.07744

Gruber, T. R. (2003). *A translation approach to portable ontology specifications. Knowledge* Acquisition, 5(2), 199–220. https://doi.org/10.1006/knac.1993.1008

Humm, P., Archer, P., Bense, H., Bernier, C., Goetz, C., Hoppe, T., Schumann, F., Siegel, M., Wenning, R., & Zender, A. (2022). *New directions for applied knowledge-based AI and machine* learning. Informatik Spektrum, 45(5), 468–482. https://doi.org/10.1007/s00287-022-01458-3

Ingrassia, S., Jacques, J., & Yao, W. (2022). *Special issue on "Models and learning for clustering and classification*." Advances in Data Analysis and Classification, 16(2), 231–234. https://doi.org/10.1007/s11634-021-00450-4

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st *International Conference on Neural Information Processing Systems* (pp. 3149–3157). https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848d3d0e6ebadc63a3-Paper.pdf

MacQueen, J. (2007). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
Mungall, C. (2024). Dynamic retrieval augmented generation of ontologies using artificial intelligence. *Journal of Biomedical Semantics*, 15(1), 19. https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-024-00156-1
Pearson, K. (2001). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572. https://doi.org/10.1080/14786440109462720
SNOMED International. (2021). SNOMED CT: *The global language of healthcare*. Retrieved from https://www.snomed.org/