SIADS 696 Milestone II Project Report

Campaign Effectiveness and Subscription Forecasting

Naz Uremek (nazliu@umich.edu), Matthew Tran (mgtran@umich.edu), Shiyi Song (sshiyi@umich.edu)

Introduction

In today's digital era, marketing has considerably evolved in data availability and technology. From ad hoc reporting insights to AI-powered machine learning methods, data driven decisions have become an integral part of marketing in various industries. Two popular themes include **campaign effectiveness** and **subscription forecasting**. Banking in particular is spearheading the use of machine learning with marketing to engage with their customers. Fine tuning the campaign process by targeting groups of individuals who are more likely to subscribe to a product allows for banks to allocate their marketing resources and budget more efficiently.

Before choosing a supervised model for a subscription campaign, the goal of the campaign is important in determining whether precision or recall of the model will be more valuable. Precision is ideal when the cost of false positives (incorrectly predicting a subscription) is high, whereas recall is ideal when it is critical for the campaign to not to miss any potential subscriptions. Our dataset consists of a bank subscription campaign targeting a small, select group of high-value customers to subscribe to the loan. In this case, we chose logistic regression which resulted in high precision with lower recall compared to decision tree classifier which resulted in high recall but lower precision.

Due to the large number of features available, this project aims to couple logistic regression with principal component analysis (PCA) to reduce dimensionality and k-means in order to cluster similar features to segment groups of customers into bank loyalist tiers. We then determine which cluster(s) is/are likely to convert to subscriptions, and similarities of features within the cluster such as demographics vs socioeconomic.

Related work

Previous research has integrated various classifier and regression methodologies. A telemarketing campaign study A Machine Learning Framework towards Bank Telemarketing Prediction discusses how effective a class membership-based (CMB) classifier is in making business decisions. The CMB model processes heterogeneous data by transforming separately each type of feature (numerical, Boolean, scaled and nominal), then uses a hybrid technique to replace missing values. Then, the most significant features are selected using AUC and f-measure. This project challenges the CMB model with k-means for classifying customers using bank membership tier clustering to determine campaign effectiveness on subscriptions.

Another study Predicting Bank Term Deposit Subscription with Logistic Regression discusses logistic regression analysis of factors contributing to bank deposit subscription and direct marketing success. This

project aims to supplement logistic regression with unsupervised learning techniques for feature creation to be used in the final predictive model.

Taking logistic regression a step further with dimensionality reduction, the study Bank Marketing Analysis Project conducted a detailed analysis to predict the success of bank telemarketing campaigns using logistic regression and other models. It emphasized the importance of understanding the factors influencing the success rates of marketing campaigns to optimize bank resources. The study combined logistic regression with PCA for dimensionality reduction and k-means clustering to segment customer groups, aiming to predict and enhance campaign effectiveness through targeted marketing strategies. The research demonstrated the application of machine learning techniques in improving the efficiency of marketing strategies in the banking sector.

Data Source, Scope, Processing

The Bank Marketing Campaign Subscriptions contains information about marketing campaigns that were conducted via phone calls from a Portuguese banking institution to their clients (total 42k records in CSV format, 6MB in size). Purpose of these campaigns is to prompt their clients to subscribe for a specific financial product of the bank called term deposit. After each call was conducted, the client had to inform the institution about their intention of either subscribing to the product (indicating a successful campaign) or not (unsuccessful campaign).

The features of interest consist of a combination of four groupings:

1. **Demographics**: Age, marital status, job, education
2. **Loan history**: Personal, home
3. **Campaign performances**: Including current and previous campaigns metrics like call duration, number of contacts since previous campaign, days since previous campaign
4. **Socioeconomic status:** Employment variance, consumer price index, consumer confidence index, Euribor 3 month interest rate, and number of employees
    - Euribor is an overnight interbank rate composed of the average interest rates from a panel of large European banks that are used for lending to one another in euros.

See Appendix B for the full data schema.

Feature Engineering

The feature engineering pipeline included following major steps:

- **One-Hot Encoding:** Convert categorical variables into a series of binary columns. See Appendix D for categorical to numeric mapping by feature.
- **Standard Scaler:** Subtract the mean and divide by the standard deviation. This standardizes features by removing the mean and scaling to unit variance.
- **Binning:** Recategorize values in groupings to reduce count of distinct values within a feature.

High Level Trends

As an exploratory phase to our project, we wanted to see trends in our dataset that gives us an idea of attributes of the bank customers, as well as potential outliers to take note of. The bar charts also give us an idea how to bin the data due to how far spread apart the data is. For example, more than half of the customers are contacted at most two times, but some have been called over 30 times. As a result, the binning was adjusted accordingly to normalize and scale the data.

Demographics

Among the clientele at the bank, the majority are 31-40 years of age (40%), married (61%), and working in administrative or blue-collar jobs (48%) with either a high school or university degree (53%).

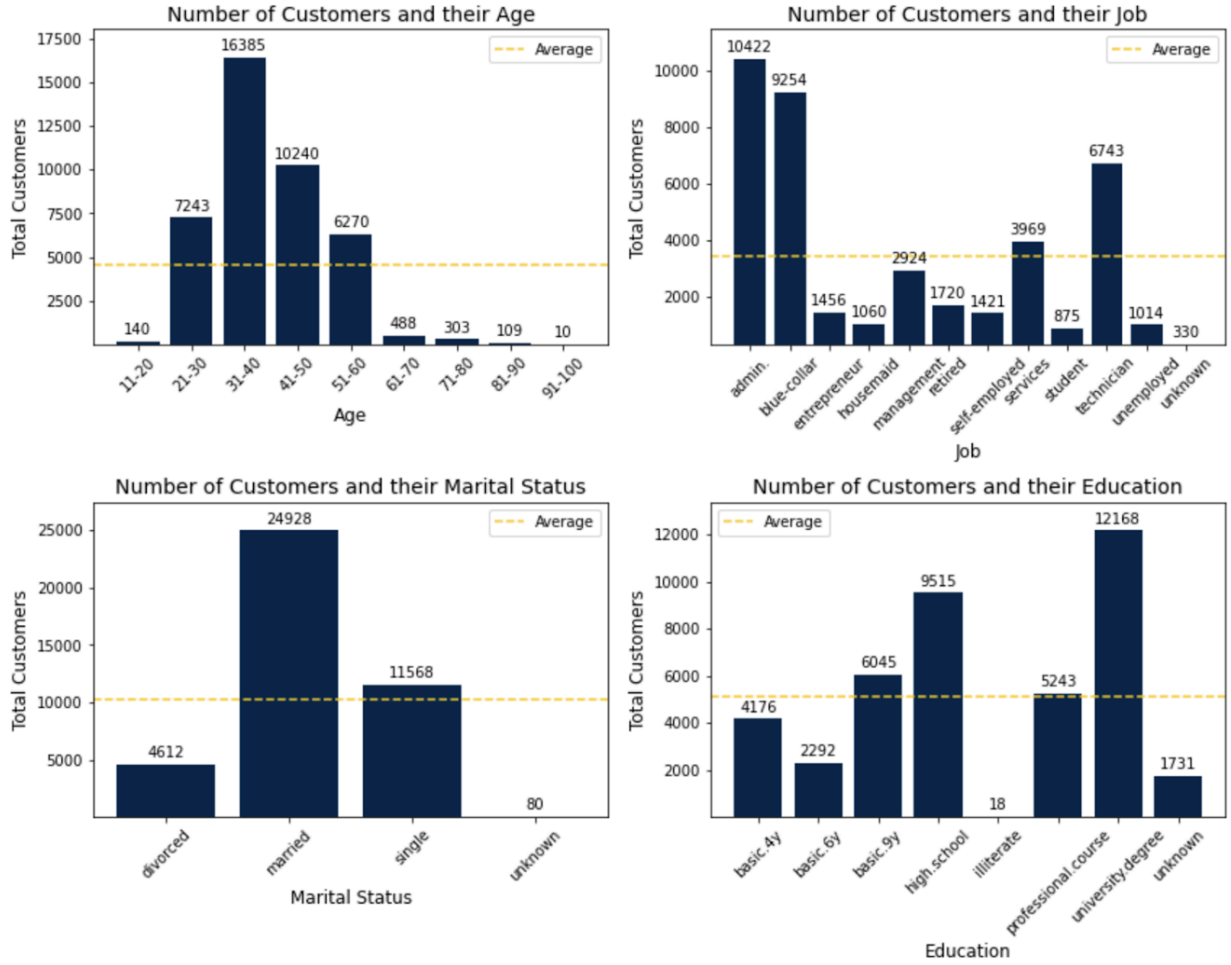


Figure 1: Customer demographics including age, job, marital status, and education

Loan history

Of the existing bank customers, more than half have a housing loan, and roughly 16% have a personal loan. Around 3% of the customers haven't provided any loan information.

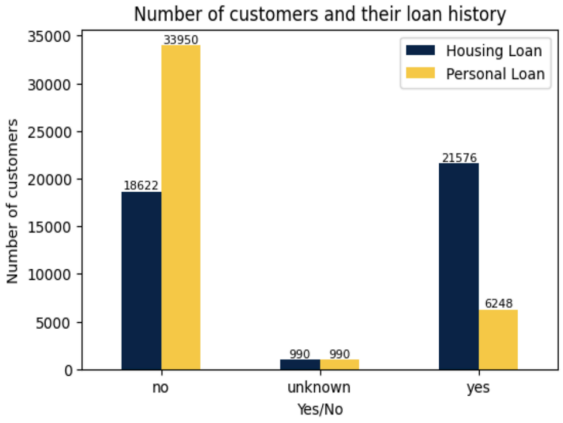


Figure 2: Customer loan history

Campaign Performance

More than half of call durations with customers were less than 200 seconds, with the highest being between 100 and 200 seconds (31%). Half the customers were also contacted at most twice. The majority of customers (86%) had no previous contacts from a campaign, and if they did, they were contacted at least 30 days ago.

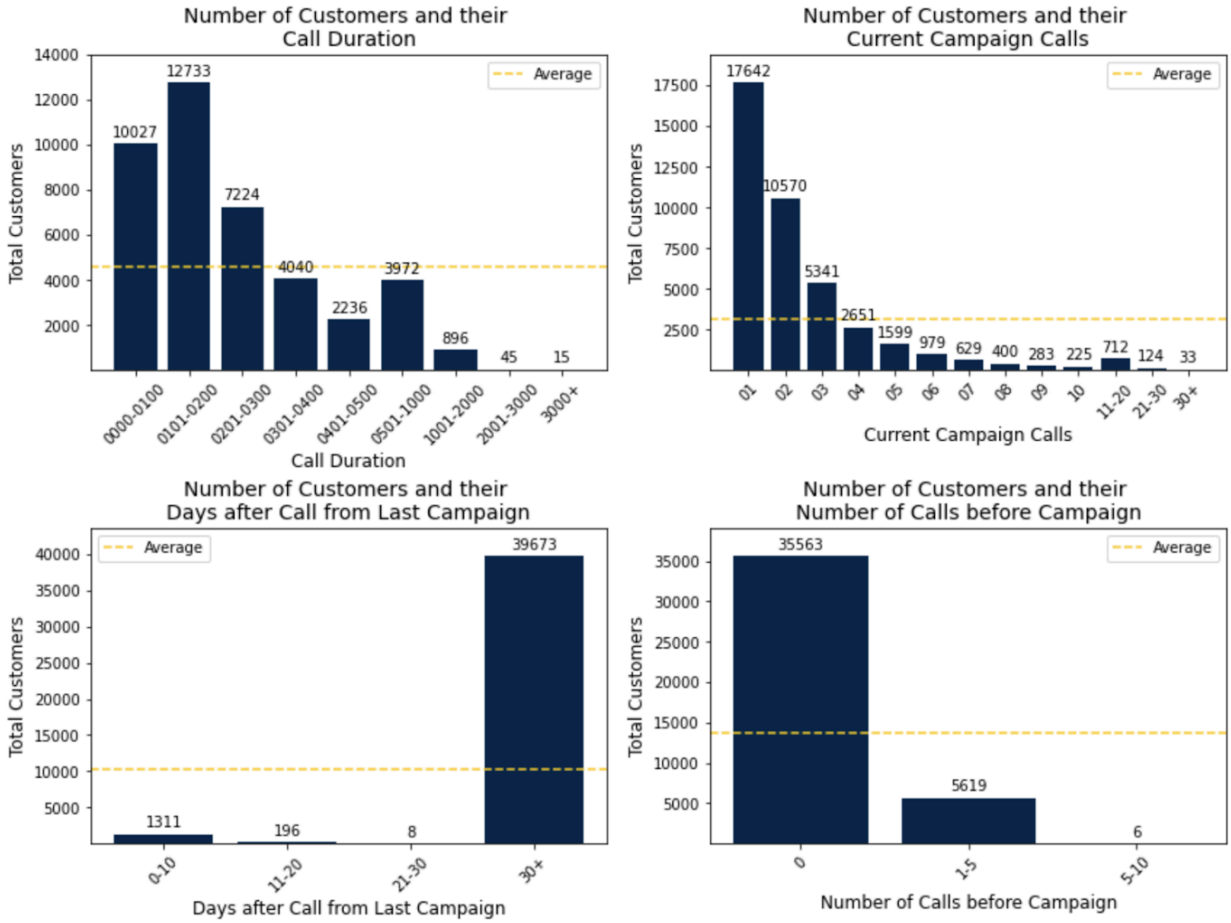


Figure 3: Customer campaign trends including call duration, total calls, days since last campaign, and previous campaign calls

Unsupervised Learning Methods

Motivation

Because the data set consists of mixed high dimensional data, we want to apply dimensionality reduction in order to reduce dimensionality and convert all variables to a numeric type. After conversion, we want to apply clustering to see if we can identify customer segments to specifically target in future marketing campaigns. Ultimately, the goal is to:

1. Reduce dimensionality
2. Cluster the principal components
3. Create a new feature called cluster membership to aid in predicting subscription

We hope to use the results of the unsupervised learning phase to aid in the supervised learning portion of this report.

Data Source

Please refer to the Data Source, Scope, and Processing section above for the details on data sources and features used. Original data schema is provided in Appendix B, post PCA schema is in Appendix C.

Unsupervised Learning Methodology

After one hot encoding and scaling, we are left with 74 features. Due to the high number of features, we decided to use Principal Component analysis as it is a popular method of dimensionality reduction. PCA will allow the extraction of features called principal components that capture the majority of the variance in the data. This reduction of features through the use of PCA can be used as a preprocessing step for supervised learning models such as Logistic Regression as it improves computational and predictive performance. Models like logistic regression can overfit as a result of high dimensional data. HIgh dimensional data also can greatly increase training time on models.

When running PCA on the data, our goal is to just reduce dimensionality without sacrificing performance in the supervised learning portion. To determine the number of principal components to use with future predictive performance in mind, we decided to use a threshold of 95% cumulative explained variance to determine the correct number of components being 29.

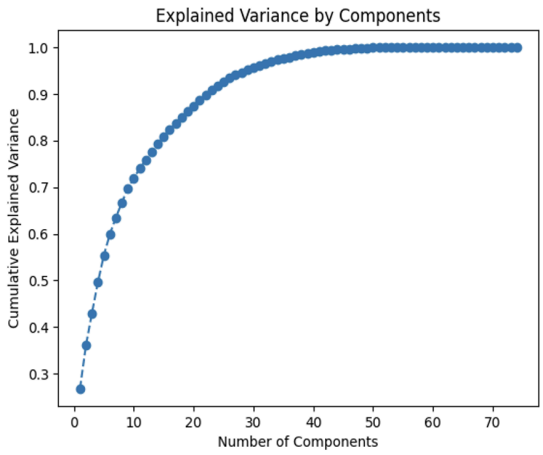
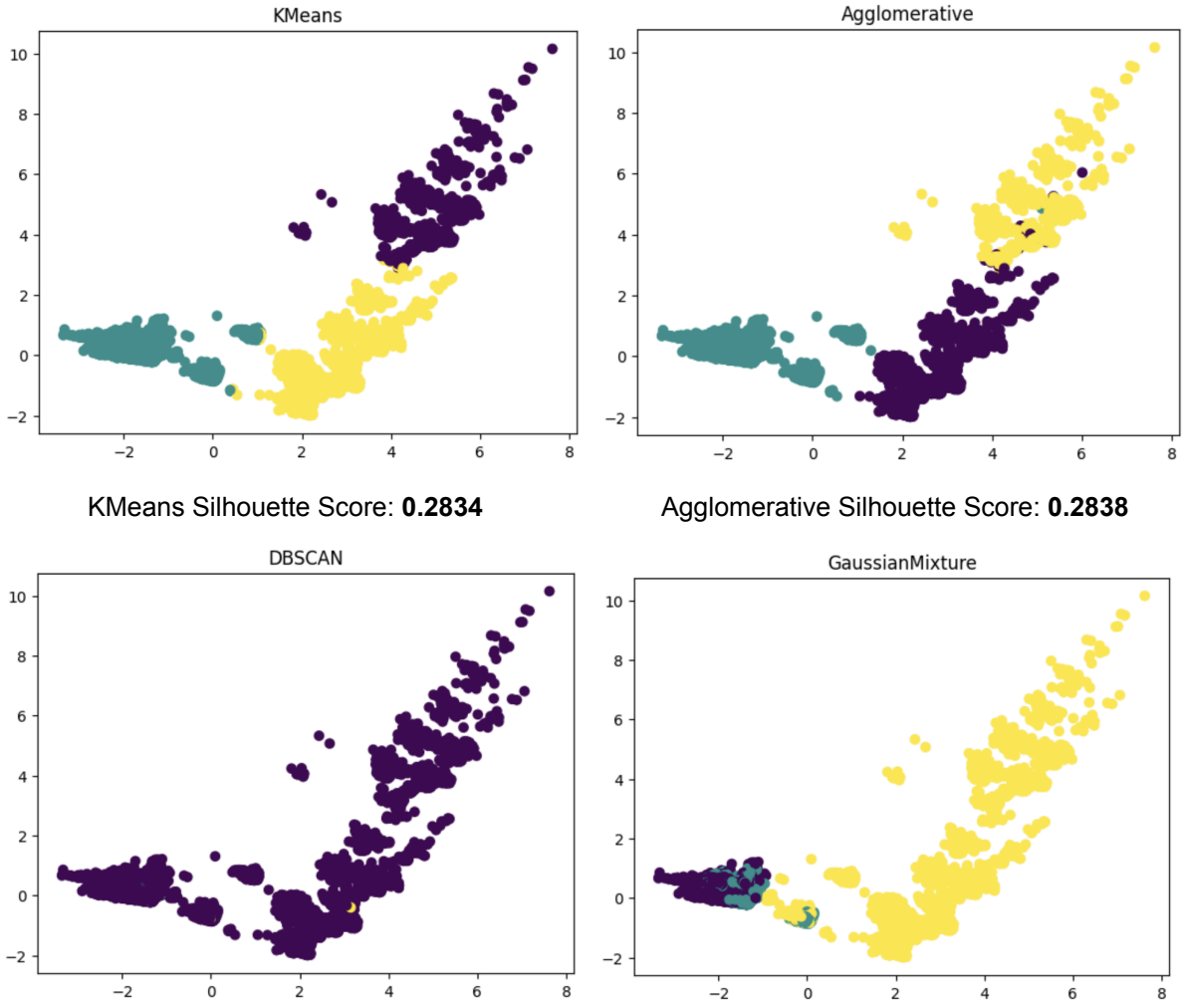


Figure 4: Cumulative explained variance by components for PCA

Another purpose of using PCA was to obtain continuous numeric features to be used in a clustering algorithm to perhaps find customer segments that could be used as potential predictors of term deposit subscription. Post PCA, to further evaluate which algorithm is the appropriate clustering method to use, we compared k-means across three other methods: DBSCAN, Agglomerative, and Gaussian Mixture. As a result, K-Means and Agglomerative silhouette scores are comparable, 0.2834 and 0.2838, respectively.

DBSCAN and GaussianMixture performed the lowest, -0.4350 and 0.0636, respectively. Because K-means and Agglomerative results have similar silhouette scores, we proceeded to continue with K-Means.



KMeans Silhouette Score: **0.2834**

Agglomerative Silhouette Score: **0.2838**

DBSCAN Silhouette Score: -0.4350

GaussianMixture Silhouette Score: 0.0636

Figure 5: Comparison of k-means with agglomerative, DBSCAN, and gaussian mixture clustering models

The k-means algorithm would be trained on the 29 features (principal components). To determine the number of clusters to use, we decided to use the elbow method and Davies-Bouldin index. This resulted in us finding an optimal cluster number of 3.

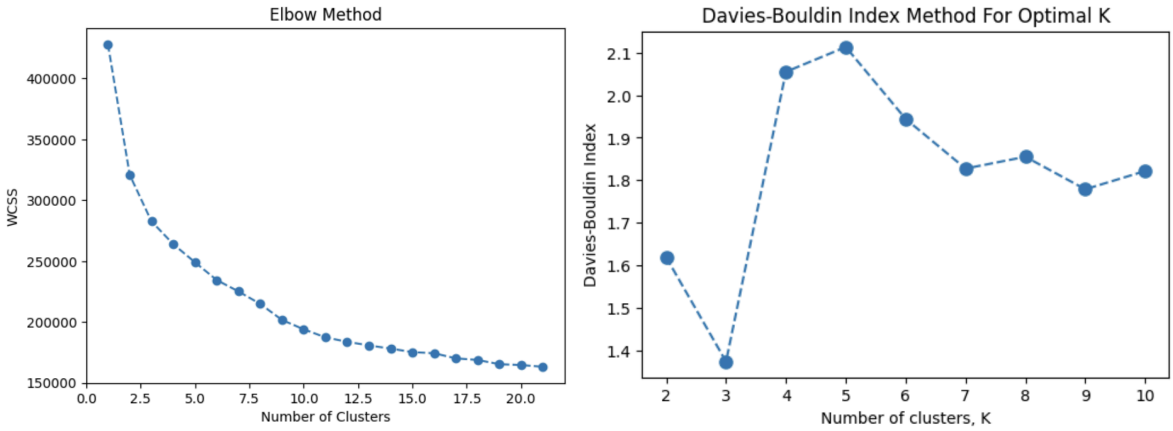


Figure 6: The use of elbow and Davies-Bouldin Index methods to identify optimal number of clusters

Results and Evaluation

Ultimately, the goal of PCA and k-means clustering was to reduce dimensionality and create a new feature 'cluster membership' to be used in the final predictive model. To determine the effectiveness of the PCA and clustering, we compared 3 models: a logistic regression trained on just one-encoded data, a logistic regression trained on one-hot encoded and scaled data, and a logistic regression trained on the principal components and the clustering membership

Features	Accuracy	Recall	Precision
Ohe only	0.9117	0.4099	0.6765
OHE & scaled	0.9125	0.4344	0.6707
PCA & Cluster Membership	0.9117	0.4214	0.6708

Model performance does not significantly change when training the logistic regression on just scaled and one hot encoded compared to a model trained on principal components and cluster membership. This is good as we were able to reduce the number of features without sacrificing predictive performance. The tradeoff is that interpretability is lowered as a result. Depending on the nature of the project, there are justifications for using either. If there was a large amount of data to process and train, using PCA would greatly reduce training time for logistic regression. If there is a small amount of training data, using just one hot encoded and scaled data would be better as you don't lose interpretability and there are no substantial computing costs. For future modeling, we will use logistic regression trained on just the one-hot encoded and scaled data to improve interpretability as we are not dealing with millions of rows of data. Including cluster membership does not seem to improve the predictive performance of the model so we may need to revise the clustering portion if we were trying to use cluster membership as a predictive indicator for subscription. We will test the predictive power of the clustering membership in the sensitivity analysis by comparing the models performance trained with and without cluster membership.

Sensitivity analysis

By removing the clustering membership from training and comparing it to a model trained on both clustering and principal components, we find that there is essentially no loss or gain in performance. This means that our cluster membership does nothing to aid in subscription prediction.

Features	Accuracy	Recall	Precision
Cluster	0.9117	0.4214	0.6708
No Cluster	0.9117	0.4229	0.6700

Key Findings:

1. The utilization of PCA is heavily reliant on the objective of the model. For large scale data, utilizing it can greatly reduce computational training data with no effect on predictive power at the cost of interpretability.
2. Clustering was not effective in creating a feature used in the final predictive model. This may be due to the PCA and should be further explored by attempting to cluster with specific features instead of principal components.

Supervised Learning Methods

Motivation

The ultimate goal of this project is to be able predict term deposit subscriptions from a bank marketing campaign. In the evaluation of the predictive performance of our logistic regression model, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) have been instrumental. As depicted in the ROC curve above, the model demonstrates an exceptional ability to distinguish between the classes. The curve reaches close to the upper left corner of the plot, indicating a high true positive rate with a low false positive rate, which is ideal in predictive performance.

Data Source

Please refer to the Data Source, Scope, and Processing section above for the details on data sources and features used. Original data schema is provided in Appendix B, post PCA schema is in Appendix C.

Results and Evaluation

We selected four metrics as the model performance measurements: accuracy, precision, recall, and f1 score. Since we're using three models to compare, we're looking to evaluate the overall performance rather than a single metric, before choosing which model to use. The ± differences indicate deviation from the metric.

For model training, a train/test split of 70/30 was used. 12 categorical features were one-hot encoded and standardized using standard scalar in order to fit into the three models (Logistic Regression,

RandomForest Classifier, and DecisionTree Classifier). In addition, we used 5-fold cross validation over the training dataset for hyperparameter tuning.

When comparing the three models, Logistic Regression resulted in higher overall accuracy, precision, and f1. Decision Tree Classifier resulted in the highest recall. RandomForest Classifier was the least performant.

Model	Accuracy	Recall	Precision	F1
Logistic Regression	**0.90968 +- 0.0036**	0.40609 +- .0229	**0.66295 +- 0.01599**	**0.50311 +- 0.01671**
RandomForest Classifier	0.90652 +- .00332	0.41044 +- 0.01548	0.63188 +- 0.01469	0.49742 +- 0.01279
DecisionTree Classifier	0.88055 +- .0026	**0.47626 +- 0.01438**	0.47073 +- 0.01152	0.47329 +- 0.00914

Based on the results, we proceeded using the logistic regression. The AUC achieved by our model is 0.94, which is a robust indicator of the model's accuracy. An AUC score close to 1 signifies that the model has a high probability of correctly classifying the inputs as positive or negative. This high score underscores the effectiveness of the preprocessing steps, feature engineering, and the logistic regression algorithm in dealing with our dataset. These results not only validate the chosen modeling approach but also inspire confidence in the potential application of this model to other similar datasets or broader contexts within the field.

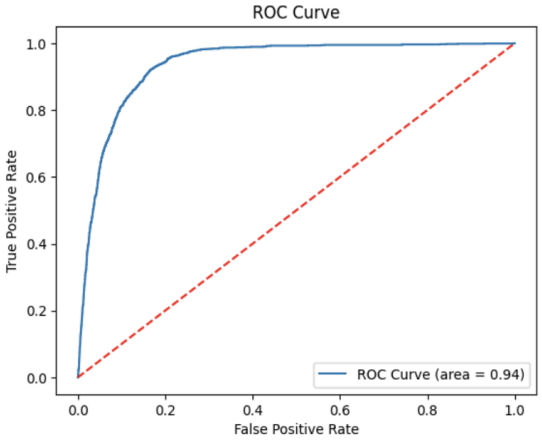


Figure 7: ROC curve from our logistic regression model

In addition to the ROC curve, we used a confusion matrix to visualize precision, recall, and accuracy. The model resulted in 43.5% recall, 67.0% precision, and 91.3% accuracy.

Sensitivity analysis

In order to choose the appropriate parameters for the logistic regression model, we applied a grid search, which resulted in best parameters C=10 and penalty=l2. C indicates the regularization strength and penalty 'l2' means that the sum of the squares of the parameters were used (Ridge Regression).

C	AUC	Recall	Precision	Accuracy
0.001	0.92974	0.30548	0.68608	0.90629
0.01	0.93591	0.40634	0.6787	0.91171
0.1	0.93718	0.43084	0.67647	0.91292
1	0.93716	0.43444	0.67075	0.91252
10	**0.93704**	**0.43516**	**0.67037**	**0.91252**
100	0.93702	0.43588	0.67073	0.9126

Key Findings:

1. The predictive model's high AUC value reinforces its utility in operational settings, suggesting that it could serve as a dependable tool in strategic decision-making processes.
2. Logistic Regression is the best for accuracy, precision, and f1 while Decision Tree works the best with recall. Since we don't know more details about the campaign, deciding whether to use a decision tree or logistic regression is unclear. Precision would be better if sending marketing materials or offers to non-subscribers would incur significant costs or negative user experience, high precision ensures that most predicted subscriptions are indeed genuine. Use recall when it's critical not to miss any potential subscriptions. For instance, if missing out on a subscription (false negative) is more costly than incorrectly predicting one (false positive), maximizing recall helps capture as many true subscriptions as possible.
3. The sensitivity analysis allows us to pick the hyperparameters resulting in the highest performing predictive model. Having regularization strength and ridge regression for penalty explicitly defined allows us to fine tune our model, enhancing the model's predictive power.

Feature Analysis

In order to understand which features strongly impact the logistic regression model, SHAP beeswarm and waterfall plots visually represent the SHAP value, which is the magnitude of each feature on its effect on the model.

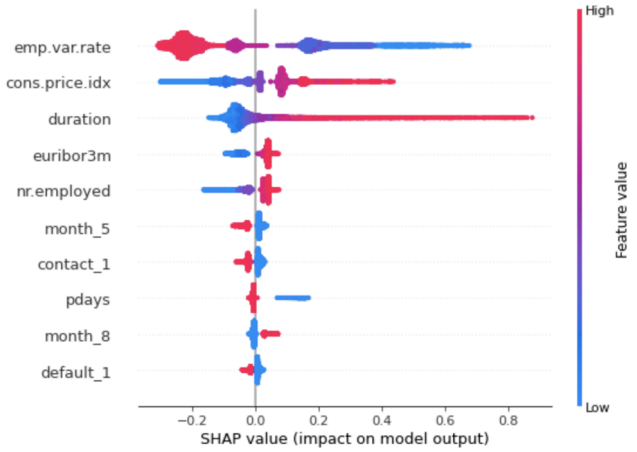


Figure 8: SHAP beeswarm plot representing top down most impactful features

According to the **beeswarm** plot, the top five impactful features for prediction are predominantly socioeconomic factors, such as employment variation rate (emp.var.rate), consumer price index (cons.price.idx), euribor 3 month interest rate (euribor3m), and number of employees (nr.employed).

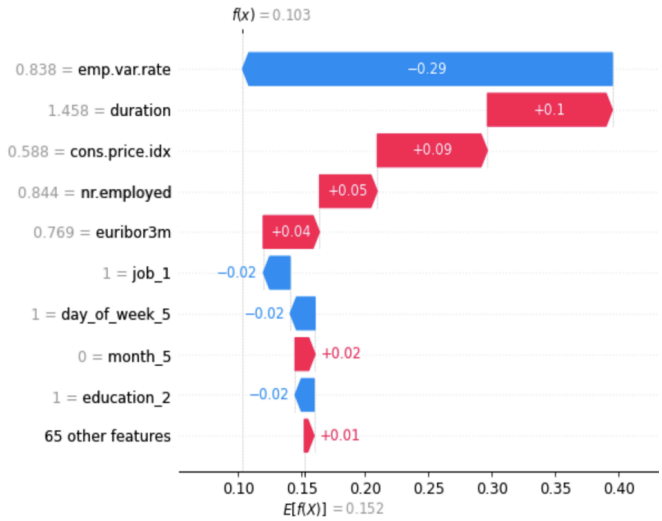


Figure 9: SHAP waterfall plot representing top down most impactful features

The one non socioeconomic feature at the top is a campaign performance feature, campaign call duration (duration).

Based on the **waterfall** plot, we can see the magnitude of impact for the top five features. Employment variation rate (emp.var.rate) had the highest negative impact at -0.29. The other features had positive impacts, with +0.1 for call duration (duration), +0.09 for consumer price index (cons.price.idx), +0.05 for euribor 3 month interest rate (euribor3m), and +0.04 for number of employees (nr.employed).

In conclusion, socioeconomic indicators have strong feature values in predicting whether or not a consumer will respond to a subscription campaign and subscribe to a product.

Taking the results from SHAP which indicate mostly socioeconomic factors impact the model's performance, also conducted **Feature Ablation** comparing two groups of features, demographics and socioeconomic features. The accuracies and standard deviations were collected using 5-fold cross-validation post logistic regression.

Demographics

model	accuracy	recall	precision	f1
non ablated features	0.912 +/- 0.004	0.442 +/- 0.017	0.666 +/- 0.028	0.531 +/- 0.008
ablated features	0.912 +/- 0.004	0.429 +/- 0.012	0.665 +/- 0.031	0.521 +/- 0.016

Socioeconomics

model	accuracy	recall	precision	f1
non ablated features	0.912 +/- 0.004	0.442 +/- 0.017	0.666 +/- 0.028	0.531 +/- 0.008
ablated features	0.909 +/- 0.003	0.394 +/- 0.019	0.663 +/- 0.027	0.493 +/- 0.015

Key Findings:

1. The SHAP analysis of the Logistic Regression study shows that the top two most important features were employment variation rate and consumer price index. Both of these features are socioeconomic indicators which show that socioeconomic factors play the largest role in determining subscription to term deposits. Term deposits are a fixed term investment that locks the deposit for a specific period of time in exchange for a higher interest rate than a more liquid high yield savings account.
2. Factors like employment variation and consumer price index can affect a potential customer's desire for more liquidity. For example, if employment variation is low (meaning decreases in employment levels) people will be more strapped for accessible money which in turn reduces subscription probability. The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. If the CPI is high, the costs of consumer goods are higher and people may want more accessible cash on hand to afford it. The goal of this ablation study is to see the effect of features involving socioeconomic factors on model performance.

Failure Analysis

Model Performance Analysis

The logistic regression model resulted in the lowest of the three models for recall (0.40609 +- .0229), indicating it misses a lot of potential subscriptions from customers (false negatives). Depending on the goal of the subscription campaign, this might not be an error, for example if the goal is to target a smaller group of customers expecting a high subscription in return. However, if the goal is to target broadly as

many customers as possible, then a model higher in recall such as decision tree would be a recommended method to test.

Error Analysis

When comparing true labels to predicted, roughly 9% of the total records are mislabeled. The top three features we can start investigating are employer variable rate (0.838), duration (1.458), and consumer price index (0.5879). This could be due to various reasons, such as insufficient training data and improper feature selection. Next steps would be to investigate these features if they cause any data leakage such as train-test contamination.

Figure 10: SHAP force plot representing the top most impactful features

Discussion

Unsupervised

Unsupervised learning may not have been effective for this use case. The goal with PCA was dimensionality reduction and we were able to achieve that by reducing the amount of features to 29. The issue with this is that it offered a similar performance on recall, precision, and accuracy as compared to just using the scaled and one-hot encoded features. Although we were able to get similar performance with less features, we sacrificed interpretability which in turn may make it more difficult to get the buy-in of non-technical stakeholders. In this particular instance it would be better to just preprocess the data and run the features directly though the logistic regression model. This does not mean PCA and clustering should never be used with this data. If there were millions of rows and this logistic regression model was used for predictions frequently, it may be beneficial to use PCA before inputting it into the LR. Although PCA is initially computationally expensive but will lead to faster training and predictions for logistic regression. The initial computational expense of using PCA can be justified by later efficiency gains when dealing with large-scale data.

Supervised

When comparing the three models (logistic regression, random forest classifier, and decision tree classifier), the top two models in respect to AUC and ROC scores were decision tree classifier and logistic regression. Logistic regression with higher precision versus decision tree classifier resulted in a higher recall. Depending on the goal of the campaign, either precision or recall could be invaluable.

When choosing **precision**, the cost of false positives (incorrectly predicting a subscription) is high. For example, precision is best when the bank subscription campaign is for a specific high-end, expensive loan, where the goal is to convert a small, select group of high-value customers to subscribe to the loan.

In our dataset, the bank was targeting customers for term deposits, which depending on campaign message, could be a high deposit required from the customer.

When choosing **recall**, it is critical for the campaign to not to miss any potential subscriptions. For banks, if the goal of the campaign is to convert as many existing customers as possible with no loan history to a specific loan, recall could help with the below performances:

- Maximizing Conversion Opportunities: Reach out to as many customers as possible. Missing a potential subscriber (false negative) means losing a potential revenue opportunity.
- Follow-Up Campaigns: Customers who are incorrectly identified as potential subscribers (false positives) might still be valuable targets for future campaigns. Reaching out to them now could increase their likelihood of conversion later.

Ethical Issues and Limitations

Based on our findings from unsupervised and supervised learnings, we conclude that bias and fairness, privacy and data security, and social and economic impact are the top ethical issues to consider.

Bias and Fairness

Models can learn and perpetuate biases present in the training data, leading to unfair treatment of certain groups. The dataset consisted of two limitations that potentially increases bias:

- **Limited data from one bank:** The subscription campaign dataset isn't diverse since it was only conducted by a Portuguese bank, mainly based off of a single Portuguese bank, so our findings may be specific to banks in Portugal. To avoid such bias, we recommend running the model using diverse and globally representative datasets.
- **Selection (voluntary) bias** may be another limitation to how the data is gathered. Oftentimes, banks request if a user would like to have their marketing information shared. Given the nature of bank marketing datasets being consent based, there would need to be additional analysis identifying the percentage of bank customers who consented, rigorous maintenance of new and historical customer data, and randomizing the sample set to represent the population as best as possible.

Privacy and Data Security

Bank data are highly sensitive and often limited to how much can be shared for analysis. Since we don't need customer details besides their demographics, loan history, campaign performance, and socioeconomic backgrounds, this ethical concern doesn't impact our model nor do we need additional processing to obfuscate any sensitive information.

- **Personal identifiable information (PII)** such as first name, last name, address, and email are limited or need to have information obfuscated. Our bank dataset successfully has PII removed and only provided demographics, previous loans with the bank, previous campaign performance, and socioeconomic factors remain.
- **Data privacy** is also vital to protect a customer's consumer behavior. Since this is a marketing campaign in a European country, GDPR would be in consideration for consent in order to share data regarding any product purchase or subscription.

Economic Impact

The goal of the model is to better understand how we can cluster bank customers based on the aforementioned four groups of attributes. Assuming our clustering successfully segments customers into low to high loyalist tiers, the output from the model would provide a list of customers to contact for a future subscription campaign. Ideally, these are high tier loyalist customers who are likely to subscribe to a given product or loan. There is a direct monetary impact of the results for the customer, and potentially for the economy (i.e. customer subscribing and campaign cost for the bank). Thus, it's important to consider the precision of the model to reduce false positives (i.e. customers who shouldn't subscribe to a loan) to avoid debt for both the customer and bank.

Statement of Work

Nazli Uremek	Matthew Tran	Shiyi Song
Introduction, Previous Research, High level visuals, Label Encoding, Davies-Bouldin Index, Clustering Model Comparison, Ethical Issues, Report Writing, SHAP Feature Analysis, Appendix	Previous Research, One-Hot Encoding, Standard Scalar, PCA, K-Means, Elbow Curve, Sensitivity analysis for both sections, supervised learning hyper- parameter tuning, Ablation Feature Analysis, Report Writing	Previous Research, ROC, AUC, Logistic Regression, Feature Analysis, Report Writing

Appendix

Appendix A - References

[1] Bracy, D., Nguyen, H. et all. (April 18, 2020). “Bank Marketing Analysis Project.” Github. https://hnguye01.github.io/DS6372/Stats2Pr2.pdf. Accessed June 2024.

[2] Ekeany. (2021). “Clustering-Mixed-Data/Notebooks/.” Github. https://github.com/Ekeany/Clustering-Mixed-Data/tree/main/Notebooks. Accessed June 2024.

[3] Keany, E. (November 1, 2021). “The Ultimate Guide for Clustering Mixed Data.” Analytics Vidhya. https://medium.com/analytics-vidhya/the-ultimate-guide-for-clustering-mixed-data-1eefa0b4743b. Accessed June 2024.

[4] Kinskey, I., Rasmus-Vorrath, J. et al. (August 18, 2017). “Predicting Bank Marketing Campaign Success Using Logistic Regression with Feature Selection and Cross Validation.” https://www.academia.edu/34329045/Predicting_Bank_Term_Deposit_Subscription_with_Logistic_Regression. Accessed May 2024.

[5] Postance, B. (April 19, 2021). “A guide to clustering large datasets with mixed data-types.” Github. https://bpostance.github.io/posts/clustering-mixed-data/. Accessed June 2024.

[6] Tekouabou, S., Gherghina,S. et al. (June 6, 2022) “A Machine Learning Framework towards Bank Telemarketing Prediction.” J. Risk Financial Manag. 2022, 15(6), 269; https://doi.org/10.3390/jrfm15060269. Accessed May 2024.

Appendix B - Original Data Schema

The dataset has around 42K records with just under 6MB total.

Data source download: Bank Marketing Campaign Subscriptions

Attribute	Description	Datatype	Example
age	age	integer	56
job	type of job (e.g. 'admin', 'technician', 'unemployed', etc)	string	housemaid
marital	marital status ('married', 'single', 'divorced', 'unknown')	string	married
education	level of education ('basic.4y', 'high.school', 'basic.6y', 'basic.9y','professional.course', 'unknown','university.degree','illiterate')	string	basic.4y
default	if the client has credit in default ('no', 'unknown', 'yes')	boolean	no
housing	if the client has a housing loan ('no', 'unknown', 'yes')	boolean	no
loan	if the client has a personal loan ? ('no', 'unknown', 'yes')	boolean	no
contact	type of communication ('telephone', 'cellular')	string	telephone
month	month of last contact	string	may
day_of_week	day of last contact	integer	mon
duration	call duration (in seconds)	integer	261
campaign	number of contacts performed during this campaign and for this client	integer	1
pdays	number of days passed by after the client was last contacted from a previous campaign	integer	999
previous	number of contacts performed before this campaign and for this client	integer	0
poutcome	outcome of previous marketing campaign ('nonexistent', 'failure', 'success')	string	nonexistent
emp.var.rate	employment variation rate - quarterly indicator	float	1.1
cons.price.idx	consumer price index - monthly indicator	float	93.994
cons.conf.idx	consumer confidence index - monthly indicator	float	-36.4
euribor3m	euribor 3 month interest rate - daily indicator	float	4.857
nr.employed	number of employees - quarterly indicator	float	5191.0
subscribed	the target variable, indicating if the client subscribed to the product ('yes') or not ('no')	boolean	no

Appendix C - Post PCA Data Schema

Dataset after clusters are created using PCA and K-Means. Our read-only version of our code can be accessed HERE.

Attribute	Description	Datatype	Example
age	age	integer	56
job	type of job (e.g. 'admin', 'technician', 'unemployed', etc)	string	housemaid
marital	marital status ('married', 'single', 'divorced', 'unknown')	string	married
education	level of education ('basic.4y', 'high.school', 'basic.6y', 'basic.9y','professional.course', 'unknown','university.degree','illiterate')	string	basic.4y
default	if the client has credit in default ('no', 'unknown', 'yes')	boolean	no
housing	if the client has a housing loan ('no', 'unknown', 'yes')	boolean	no
loan	if the client has a personal loan ? ('no', 'unknown', 'yes')	boolean	no
contact	type of communication ('telephone', 'cellular')	string	telephone
month	month of last contact	string	may
day_of_week	day of last contact	integer	mon
duration	call duration (in seconds)	integer	261
campaign	number of contacts performed during this campaign and for this client	integer	1
pdays	number of days passed by after the client was last contacted from a previous campaign	integer	999
previous	number of contacts performed before this campaign and for this client	integer	0
poutcome	outcome of previous marketing campaign ('nonexistent', 'failure', 'success')	string	nonexistent
emp.var.rate	employment variation rate - quarterly indicator	float	1.1
cons.price.idx	consumer price index - monthly indicator	float	93.994
cons.conf.idx	consumer confidence index - monthly indicator	float	-36.4
euribor3m	euribor 3 month interest rate - daily indicator	float	4.857
nr.employed	number of employees - quarterly indicator	float	5191.0
Components 1-29	PCA component values (29 components total)	float	2.042440
PCA Cluster label	clusters created from k-means and PCA (0, 1, 2)	integer	1

Appendix D - Categorical to Numeric Mapping

Numerical identities of values after using one hot encoding.

Value	Age	Job	Education	Marital	Month	Day of Week
0	11-20	admin	basic.4y	divorced		
1	21-30	blue-collar	basic.6y	married	Jan	Mon
2	31-40	entrepreneur	basic.9y	single	Feb	Tue
3	41-50	housemaid	high.school	unknown	Mar	Wed
4	51-60	management	illiterate		Apr	Thu
5	61-70	retired	professional.course		May	Fri
6	71-80	self-employed	university.degree		Jun	
7		services	unknown		Jul	
8		student			Aug	
9		technician			Sept	
10		unemployed			Oct	
11		unknown			Nov	
12					Dec	

Value	Default	Housing	Loan	Contact	Outcome	Subscribed
0	no	no	no	cellular	failure	no
1	unknown	unknown	unknown	telephone	nonexistent	yes
2	yes	yes	yes		success	