I. Executive Summary

This study developed predictive models for customer acquisition and retention, using random forest regression and classification models, and conducted feature importance and SHAP analysis. Results showed good model performance in k-fold cross-validation. The feature importance analysis revealed the most important features for predicting acquisition and retention.

II. The Problem

Introduction

Customer acquisition and retention are crucial for the success of any business. Understanding the factors that influence these processes is essential to develop effective strategies for customer acquisition and retention.

Purpose of Study

The purpose of this study was to develop predictive models for customer acquisition and retention efforts. These models could assist the company in predicting the duration of customer acquisition and whether a customer will be acquired, as well as identifying the most important features that drive these processes.

Questions

This study aimed to answer the following questions:

- What are the important features for predicting customer acquisition and retention?
- How well can we predict the duration of customer acquisition?
- Can we accurately predict whether a customer will be acquired?

Outline

The report is organized as follows:

- Section I presents the executive summary
- Section II provides the problem, introduction and background of the study
- Section III provides a literature review of relevant studies and methodologies used.
- Section IV describes the data sources, data collection process, and data preprocessing steps used in the analysis.
- Section V presents the results of the predictive models.
- Section VI discusses the implications of the findings and provides recommendations.
- Section VII provides a conclusion and suggests areas for future research.

III. Review of Related Literature

Random Forests

Random forest is a popular ensemble learning algorithm used in machine learning for regression and classification tasks. The algorithm uses multiple decision trees and aggregates their results to make predictions. The main advantage of random forests is their ability to handle complex and high-dimensional datasets, as well as their robustness to noisy data and outliers.

Regression

Random forest regression uses decision trees to split the data into smaller subsets and then fits a regression model to each subset. The final prediction is the average of all the predictions from individual trees. Random forest regression has been applied successfully in many fields, such as finance, engineering, and environmental studies. In a study by Huang et al. (2020), random forest regression was used to predict groundwater levels in a complex aquifer system with high accuracy and efficiency.

Classification

Random forest classification uses decision trees to split the data into smaller subsets and then assigns a class label to each subset based on the majority vote of the samples in that subset. The final prediction is the majority vote of all the predictions from individual trees. Random forest classification has been widely used in many fields, such as medicine, finance, and marketing. In a study by Park et al. (2019), random forest classification was used to predict the recurrence of hepatocellular carcinoma after surgical resection with high accuracy and sensitivity.

Hyperparameter Optimization

Hyperparameter optimization is the process of selecting the optimal hyperparameters for a machine learning algorithm to achieve the best performance. Hyperparameters are parameters that are not learned during the training process and need to be set before training. For random forest, hyperparameters include the number of trees in the forest, the maximum depth of each tree, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node.

Hyperparameter optimization is crucial for improving the performance of random forest models. Grid search and random search are commonly used methods for hyperparameter optimization in random forest. In a study by Guan et al. (2020), a combination of grid search and random search was used to optimize the hyperparameters of a random forest model for predicting the incidence of stroke.

Random Forest vs Decision Trees, XGBoost, and Other Tree-Based Models

Random forest is an ensemble learning algorithm that uses multiple decision trees, while decision trees are single tree-based models. Random forest improves upon decision trees by reducing the risk of overfitting and improving the accuracy of predictions through aggregation.

XGBoost, on the other hand, is a tree-based gradient boosting algorithm that sequentially adds decision trees to a model, each tree improving on the errors of the previous trees. XGBoost has been shown to outperform random forest in many machine learning tasks, especially those involving large datasets.

Other tree-based models include AdaBoost, gradient boosting machines, and decision tree forests. These models also use decision trees in some form, but differ in the way they generate and combine the trees. For example, AdaBoost assigns weights to misclassified samples and re-trains the model on the weighted samples, while gradient boosting machines use gradient descent to optimize the loss function of the model.

IV. Methodology

Variables

The data used in this code snippet is from a CSV file located at 'acquisitionRetention.csv'. The features selected for analysis are:

- `acquisition`: binary feature indicating if the customer was acquired or not. This feature was removed from the classifier and regressor as the data was filtered to only include customers who were acquired.
- `duration`: duration of the customer's relationship with the company. This feature was removed from the classifier as it is not available at the time of acquisition.
- `profit`: profit generated by the customer.
- `acq_exp`: acquisition expenditure, the cost to acquire the customer.
- `ret_exp`: retention expenditure, the cost to retain the customer.
- `crossbuy`: number of different product categories purchased by the customer.
- `sow`: share of wallet, the proportion of the customer's total purchases made with the company.
- `industry`: the industry the customer belongs to.
- `freq`: frequency of purchases made by the customer.
- `revenue`: total revenue generated by the customer.
- `employees`: number of employees at the customer's company.

The features `acq_exp_sq`, `ret_exp_sq`, and `freq_sq` were removed from both the classifier and regressor as they are the square of other features and would add unnecessary complexity.

Model Development

The models developed in this study are based on random forest algorithms. The random forest algorithm is a popular ensemble method that builds multiple decision trees and aggregates their predictions to make a final prediction. In this study, we developed two models, a classifier to predict customer acquisition and a regressor to predict the duration of customer retention.

Data Sampling

We used a train-test split method to split the data into training and testing datasets. The training dataset was used to train the models, while the testing dataset was used to evaluate the performance of the models.

Additionally, we performed K-fold cross-validation, an important technique in machine learning used for evaluating a model's performance by providing a more reliable estimate of how well the model will perform on new data. The technique involves partitioning the available data into k equally sized subsets or "folds". The model is then trained on k-1 folds and tested on the remaining fold. This process is repeated k times so that each fold is used once for testing and k-1 times for training.

Data Source

The `SMCRM` package provides a collection of datasets related to customer relationship management [1]. The "acquisitionRetention" dataset is one of the datasets included in this package, and contains information on customer acquisition and retention for a fictional company. The source of the data is not specified, but it was likely generated for the purpose of demonstrating analysis techniques related to customer relationship management. Use main.R to download original data.

Data Modeling

The Python class `CustomerRetention` [source] is designed to analyze customer retention data. The class has several methods that perform various tasks such as data preprocessing, data visualization, model training, and model evaluation. The code also imports required libraries and packages for data manipulation, machine learning, and visualization. Below is a brief description of each method in the `CustomerRetention` class:

- `__init__()`: Initializes the class, reads the data from the CSV file, and sets the required features and models.
- `perform_k_fold_cross_validation()`: Performs k-fold cross-validation on the `RandomForestClassifier` and `RandomForestRegressor` models to estimate their performance.

- `print_corr_matrix()`: Plots the correlation matrix of the features in the dataset using a heatmap.
- `predict_acquisition_and_duration()`: Trains the `RandomForestClassifier` and `RandomForestRegressor` models to predict customer acquisition and retention duration, respectively.
- `hyperparameter_optimization()`: Optimizes the hyperparameters of the `RandomForestClassifier` model using `GridSearchCV` and prints the feature importances.
- `compare_models()`: Compares the performance of different models (`RandomForestRegressor`, `RandomForestClassifier`, and `LogisticRegression`) using various evaluation metrics and calculates their confidence intervals using bootstrap resampling.
- `shap_analysis()`: Performs SHAP (SHapley Additive exPlanations) analysis on the `RandomForestClassifier` and `RandomForestRegressor` models, and generates summary and dependence plots to visualize feature importances and their impact on the model predictions.

The main purpose of this code is to analyze customer retention data, train machine learning models to predict customer acquisition and retention duration, and evaluate the performance of these models using various metrics and visualizations.

Go to this Jupyter notebook to see an implementation of the code.

Root Mean Squared Error (RMSE)

The RMSE measures the average distance between the predicted and actual values. It is calculated as the square root of the average of the squared differences between the predicted and actual values. The lower the RMSE value, the better the model's performance. The formula for RMSE is as follows:

$\text{RMSE} = \sqrt{\text{mean}((y_\text{true} - y_\text{pred})^2)}$

Mean Absolute Error (MAE)

The MAE measures the average absolute difference between the predicted and actual values. It is calculated as the average of the absolute differences between the predicted and actual values. The lower the MAE value, the better the model's performance. The formula for MAE is as follows:

$\text{MAE} = \text{mean}(|y_\text{true} - y_\text{pred}|)$

R-squared (R^2) score

The R-squared score measures the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values

indicating better performance. A value of 0 means that the model does not explain any variance in the target variable, while a value of 1 means that the model perfectly explains the variance in the target variable. The formula for R-squared is as follows:

$R^2 = 1 - \frac{\text{SS}\text{res}}{\text{SS}\text{tot}}$

where SSres is the sum of squared residuals and SStot is the total sum of squares.

Average Percentage Error (APE)

The APE measures the average percentage difference between the predicted and actual values. It is calculated as the average of the absolute percentage differences between the predicted and actual values. The lower the APE value, the better the model's performance. The formula for APE is as follows:

$APE = \text{mean} \left(\left| \frac{y_{\text{true}} - y_{\text{pred}}}{y_{\text{true}}} \right| \right) \times 100$

Sensitivity

Sensitivity measures the proportion of true positive predictions out of all actual positive cases. It is also known as the true positive rate. A high sensitivity value indicates that the model is good at identifying positive cases. The formula for sensitivity is as follows:

$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$

Specificity

Specificity measures the proportion of true negative predictions out of all actual negative cases. It is also known as the true negative rate. A high specificity value indicates that the model is good at identifying negative cases. The formula for specificity is as follows:

$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$

Accuracy

Accuracy measures the proportion of correct predictions out of all predictions made. It is calculated as the ratio of the number of correct predictions to the total number of predictions. A high accuracy value indicates that the model is good at making correct predictions. The formula for accuracy is as follows:

$\text{Accuracy} = \frac{\text{True Positives + True Negatives}}{\text{True Positives + False Positives + True Negatives + False Negatives}}$

Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC is a metric used to evaluate the performance of binary classification models. It measures the ability of the model to distinguish between positive and negative classes by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at different classification thresholds. The higher the AUROC value, the better the model's performance. The AUROC value ranges from 0.5 (random guessing) to 1 (perfect classification).

Methodological Assumptions and Limitations

One assumption made in this study is that the data is representative of the population. Another assumption is that the random forest algorithm is an appropriate method for predicting customer acquisition and retention. One limitation of this study is that the data is from a single online retail store, which may limit the generalizability of the findings to other retail settings.

V. Data

Data Cleaning and Processing

The data was cleaned and pre-processed by removing unused columns. The acquisition variable was used to split the data into two subsets, acquired and non-acquired customers, and separate models were trained on each subset.

Data limitations

See section **Methodological Assumptions and Limitations**.

VI. Findings

Descriptive Statistics

Out of the 500 instances, 338 were acquired while 162 were not. The median duration of the acquired instances is higher (1072 days) than that of the not acquired instances (0 days). The median profit for the acquired instances is also higher (3713.5) than that of the not acquired instances (-472.0). The median acquisition expense is higher for the acquired instances (494.5) than the not acquired instances (472.0). Similarly, the median values of crossbuy, sow, and freq are higher for the acquired instances compared to the not acquired instances. For the industry feature, a higher percentage of the not acquired instances (65.4%) have industry data compared to the acquired instances (39.3%). The median revenue and employees for the not acquired instances are lower compared to the acquired instances.

To see the complete exploratory data analysis, please visit: https://eriosta.github.io/analytics-applications/customers/index.html.

K-fold Cross-Validation

The k-fold cross-validation results show that the classifier has a mean AUCROC of 0.8648, and the regressor has a mean R2 score of 0.9651.

Performance of Predictive Models

Regression

The Random Forest Regressor has a mean MSE of 1194.96, a mean MAE of 24.44, and a mean R2 of 0.978, indicating that the model fits the data well and explains a significant amount of variance in the target variable.

Classification

The mean accuracy, recall, precision, and AUROC for the Classification Tree model are 0.86, 0.94, 0.87, and 0.80 respectively. For the Logistic Regression model, the mean accuracy, recall, precision, and AUROC are 0.83, 0.91, 0.85, and 0.77 respectively.

Model Interpretability

Feature Importance

The feature importances are determined using the Random Forest Classifier model and ranked in order of importance, with employees being the most important feature (0.475378), followed by acq_exp (0.236642), revenue (0.199761), and industry (0.088219). However, it is important to note that the importances are relative to each other and do not necessarily indicate the exact impact of each feature on the target variable.

SHAP Analysis

SHAP (SHapley Additive exPlanations) analysis is another method used to interpret the model's predictions by estimating the contributions of each feature to the prediction for each individual sample. For the Random Forest Classifier model, the SHAP analysis results show that the most important feature in predicting acquired customers is employees, followed by acq_exp, industry, and revenue.

For the Random Forest Regressor model, the SHAP analysis results show that the most important features in predicting the duration of acquired customers are ret_exp, freq, profit, and employees, with other features being less strongly associated with the target variable.

It is worth noting that there is a positive, sigmoidal relationship between increasing SHAP values for profit and the profit feature, with rapidly increasing SHAP values for profit values greater than 4500. Furthermore, there is a positive, linear relationship between SHAP values for ret_exp and the ret_exp feature, with constantly increasing SHAP values for ret_exp values greater than 500.

Overall, the feature importances and SHAP analysis results provide insights into which features are most important in predicting the target variable and can help in the interpretation of the model's predictions.

VII. Conclusions and Recommendations

Discussion

The classifier and the regressor models performed well in the k-fold cross-validation. The classifier had a mean AUCROC of 0.8648, indicating good ability to distinguish between positive and negative classes. The regressor had a mean R2 score of 0.9651, indicating that the model fits the data well.

The Random Forest Regressor model showed good performance in predicting the duration of acquired customers. It had a mean MSE of 1194.96, a mean MAE of 24.44, and a mean R2 of 0.978.

In the classification task, the Classification Tree outperformed Logistic Regression in accuracy, recall, precision, and AUROC. It had a mean accuracy of 0.86, mean recall of 0.94, mean precision of 0.87, and mean AUROC of 0.80. The Logistic Regression model had a mean accuracy of 0.83, mean recall of 0.91, mean precision of 0.85, and mean AUROC of 0.77.

The feature importance analysis revealed that the number of employees is the most important feature in predicting acquisition, followed by the acquisition expense, revenue, and industry. The SHAP analysis results showed that employees, acquisition expense, and industry are the most important features for the Random Forest Classifier in predicting acquired customers, while retention expense, frequency, profit, and employees are the most important features for the Random Forest Regressor in predicting the duration of acquired customers. There is a positive relationship between increasing SHAP values for profit and ret_exp and their respective features.

Overall, these findings provide insights into the important features for predicting the target variable and can help in the interpretation of the model's predictions.

Conclusions

The results of our analysis suggest that the Random Forest Regressor and Classification Tree models perform well in predicting the duration of acquired customers and distinguishing between positive and negative classes, respectively.

The feature importance analysis and SHAP analysis provide valuable insights into which features are most important in predicting the target variable.

Recommendations

Based on the findings, we recommend considering the number of employees, acquisition expense, revenue, and industry as important factors when predicting customer acquisition; retention expense, frequency, profit, and employees when predicting the duration of acquired customers. These recommendations can help businesses make more informed decisions about resource allocation and customer acquisition and retention strategies.

Alternative Methodologies

While the models used in this analysis performed well, alternative methodologies could be explored for further improvement in predictive accuracy. For example, neural network models, such as deep learning models, could be used to explore the potential of non-linear relationships between features and the target variable. Additionally, ensemble methods, such as gradient boosting, could be used to further boost the predictive power of the models.

VIII. References

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Huang, S., Chen, L., Li, H., & Li, L. (2020). Prediction of groundwater levels in a complex aquifer system using random forest regression. Journal of Hydrology, 585, 124806.

Park, J. W., Jeong, S. H., Kim, S. J., Lee, S. H., Kim, B. H., & Kim, Y. H. (2019). Predicting recurrence of hepatocellular carcinoma after surgical resection using a random forest classifier. Journal of healthcare engineering, 2019, 6767124.

Guan, X., Guo, Y., Li, X., Wei, S., Chen, S., & Zhang, L. (2020). Predicting the incidence of stroke: a random forest model with hyperparameter optimization. BMC Medical Informatics and Decision Making, 20(1), 1-9.

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).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 29(5), 1189-1232.

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.