International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 11, Issue 03, March 2024
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, Email : editoraarf@gmail.com

"A Study of Predicting NPS Status through Multi-Algorithm Classification Models"

Dr.Dnyandev Nitve
HOD, Department of Commerce
Pvg's College of Science and Commerce Shivdarshan, Parvati Pune

Dr. Sanjaykumar Gaikwad
Principal- Pvg's College of Science and Commerce Shivdarshan, Parvati Pune

ABSTRACT
Customer satisfaction is a pivotal metric in today's competitive business landscape, with the Net Promoter Score (NPS) serving as a key indicator. This research paper focuses on the development and implementation of a predictive analytics solution to forecast NPS status using a diverse set of classification algorithms. The research involves comprehensive data collection, preprocessing, and exploratory data analysis to understand the intricacies of customer feedback. A variety of machine learning algorithms, including logistic regression, decision trees, random forest, support vector machines, k-nearest neighbors, and gradient boosting, will be employed to construct robust classification models. The models will undergo rigorous training, evaluation, and fine-tuning, with a keen emphasis on interpretability and ensemble methods. The chosen model will be deployed and integrated into real-world scenarios, with continuous monitoring and periodic updates to ensure sustained accuracy. This research paper seeks to offer businesses a reliable tool for predicting NPS status, facilitating proactive measures to enhance customer satisfaction and loyalty.

Objectives of the study
1. To Study the research begins with a thorough exploration of the dataset through data wrangling and exploratory data analysis (EDA). This foundational step aims to enhance data understanding and ensure the dataset's readiness for model development.
2. An analysis of the distribution of the target variable (NPS status) provides insights into the class distribution, highlighting the imbalance that necessitates the use of oversampling techniques.
3. Encoding techniques are applied to categorical variables, and the data is standardized and normalized for consistency and comparability.
4. To Study the dataset is split into training and validation sets to facilitate model training and evaluation.
5. Feature selection techniques are employed to identify and retain the most relevant features, optimizing model performance and interpretability.

Scope of the study

International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 11, Issue 03, March 2024
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, Email : editoraarf@gmail.com

1. **Comprehensive Data Exploration:** The research will extensively explore and analyse the dataset through data wrangling and exploratory data analysis (EDA) to ensure a thorough understanding of the underlying patterns and relationships.
2. **Addressing Class Imbalance:** The study explicitly considers the challenge of highly imbalanced class distribution in predicting NPS status and aims to mitigate this issue through the application of oversampling techniques.
3. **Data Pre-processing Techniques:** The research includes the implementation of data encoding, scaling (standardization and normalization), and train-validation split to prepare the dataset for effective model training and evaluation.
4. **Feature Selection:** Feature selection methods will be employed to identify and retain the most relevant features, optimizing the efficiency and interpretability of the classification models.
5. **Diverse Classification Models:** The study involves the development and comparison of multiple classification algorithms, including K-Nearest Neighbors (KNN), Decision Tree, and Random Forest, to assess their effectiveness in predicting NPS status.
6. **Model Performance Metrics:** The research evaluates the classification models without oversampling techniques using key performance metrics such as accuracy, precision, recall, and F1-score. This provides a comprehensive assessment of the models' predictive capabilities.

Research Methodology

1) Primary Data and Secondary Data

2) Sample Size

The sample size for Training the classification models is 4,989 whereas the sample size for testing the output of the classification models is 364.

Sampling Method

1. K-Nearest Neighbour (KNN)
2. Decision Tree
3. Random Forest

K- Nearest Neighbor (KNN)

K-Nearest Neighbors (KNN) is a simple machine learning algorithm for classification and regression. It classifies a data point by identifying the K nearest neighbors in the training set based on a chosen distance metric. For classification, it uses majority voting among the neighbors to assign a class, while for regression, it calculates the average of their target values. KNN's decision boundaries adapt to the data distribution, and the choice of K is a crucial parameter. It's easy to implement but can be computationally expensive for large datasets, and its performance may be affected by the curse of dimensionality.

Decision Tree

A Decision Tree is a machine learning algorithm that models decisions through a tree-like

International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 11, Issue 03, March 2024
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, Email : editoraarf@gmail.com

structure. It recursively splits data based on features, with each internal node representing a decision point and each leaf node providing a final outcome or prediction. The algorithm selects splits by optimizing criteria like Gini impurity for classification or mean squared error for regression. Decision Trees are interpretable, handle both numerical and categorical data, and implicitly perform feature selection. However, they can be prone to overfitting and may not capture complex relationships. Pruning and ensemble methods like Random Forests mitigate these issues.

Random Forest

Random Forest is an ensemble learning technique that combines multiple Decision Trees for improved accuracy and robustness. It creates a diverse set of trees by training on random subsets of the data and features. Each tree in the forest independently makes predictions, and the final output is determined by aggregating these predictions, often through voting or averaging. Random Forest mitigates overfitting observed in individual Decision Trees and enhances generalization to new data. It is widely used for classification and regression tasks, demonstrating resilience to noisy data and providing feature importance insights.

Observations of the study

Model	Train accuracy	Test accuracy	Precision macro average	Recall macro average
KNN	75	69.4	61	56
Decision tree	73.1	68.8	64	51
Random forest	91.6	70.6	65	56

1. Data Imbalance: The analysis revealed a significant class imbalance in the target variable (NPS Status). The distribution of the target classes shows a notable difference, with the "Promotor" class being predominant compared to "Detractor" and "Passive" classes.

2. Feature Analysis:

a) The notebook performs exploratory data analysis (EDA) on various features, including

International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 11, Issue 03, March 2024
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, Email : editoraarf@gmail.com

categorical variables related to patient demographics, feedback on different hospital services, and numerical variables like age, estimated cost, and length of stay.

b) Bivariate analysis is conducted to explore the relationship between different features and the target variable, providing insights into how various factors might influence the NPS status.

3. Data Pre-processing:

a) Data encoding and scaling techniques are applied to prepare the dataset for machine learning models.

b) One-hot encoding is used for categorical variables, label encoding for certain ordinal variables, and min-max scaling for numerical variables.

c) The dataset is split into training and validation sets.

4. Feature Selection:

a) The notebook addresses multicollinearity by calculating Variance Inflation Factors (VIF) and dropping highly correlated features.

b) Feature importance is explored using a Random Forest classifier, but no explicit feature filtering is performed based on importance.

5.Modelling:

a) K-Nearest Neighbour (KNN) and Decision Tree (CART) models are implemented for classification without oversampling.

b) The KNN model is tuned for the optimal number of neighbours (K) based on accuracy.

c) Hyperparameter tuning is performed for Decision Tree and Random Forest using grid search.

6) Model Evaluation:

a) The models' performance is evaluated using accuracy metrics, confusion matrices, and classification reports on both the training and validation datasets.

b) Random Forest achieves higher accuracy compared to KNN and Decision Tree on the validation set.

Limitations of the Study

1. **Assumption of Stationarity:** The study assumes that the characteristics of the dataset, particularly the relationships between features and NPS status, remain relatively stable over time. Any changes in these dynamics may impact the model's predictive accuracy.
2. **Dependency on Dataset Quality:** The effectiveness of the classification models is contingent on the quality and representativeness of the dataset. Biases, missing data, or outliers in the dataset may introduce limitations to the generalizability of the models.
3. **Algorithm Sensitivity:** The choice of classification algorithms may be sensitive to the specific characteristics of the dataset. Certain algorithms may perform better in specific scenarios, and the study acknowledges this sensitivity.
4. **Oversampling Impact:** While oversampling techniques, specifically Synthetic Minority Over-sampling Technique (SMOTE), are applied to address class imbalance, the impact of oversampling on the overall model performance will be considered, and

potential trade-offs will be discussed.

5. **External Factors:** External factors, such as changes in customer behavior or external market conditions, are beyond the scope of the project. These factors could influence the predictive accuracy of the models but are not explicitly addressed.
6. **Limited Ensemble Techniques:** The study focuses on individual classification algorithms and incorporates oversampling techniques. However, more advanced ensemble techniques may not be explored in-depth due to the scope constraints. by acknowledging these limitations, the study aims to provide a realistic and transparent assessment of the developed classification models for predicting NPS status in the context of the specified scope.

CONCLUSION

The notebook provides a comprehensive analysis of a classification problem predicting Net Promoter Score (NPS) status in a healthcare setting. It addresses data preprocessing, explores feature importance, and implements machine learning models for classification. The findings suggest potential areas for improvement, including handling class imbalance, refining feature selection, and tuning model hyperparameters.

To enhance the model's robustness and interpretability, further steps could involve experimenting with different algorithms, fine-tuning hyperparameters, and incorporating advanced techniques for handling class imbalance. Additionally, leveraging feature importance insights can guide the selection of key factors influencing the NPS status.

In conclusion, the notebook serves as a foundation for predictive modelling in the healthcare domain, with opportunities for refinement and expansion in subsequent iterations.

REFERENCES:

Research Papers:

- https://cambridgeservicealliance.eng.cam.ac.uk/system/files/documents/2016OctoberPaper_FallacyoftheNetPromoterScore.pdf
- https://www.mdpi.com/915752
- https://publications.dlpress.org/index.php/ijic/article/view/41

Websites:

- https://pages.mtu.edu/~shanem/psy5220/daily/Day13/treesforestsKNN.html
- https://www.qualtrics.com/au/experience-management/customer/net-promoter-score/
- https://www.javatpoint.com/nps-in-machine-learning