Machine Learning Project

Muhammad Uzair - 29414

Pipeline Overview:

pipeline is designed to handle classification tasks with imbalanced datasets

Data Loading and Cleaning:

You load the dataset from a given file path and handle any missing or inconsistent data.

Class Label Encoding:

You encode the class labels to numerical values, typically 0 and 1 for binary classification tasks.

Data Transformation:

You preprocess the data, including standardizing numerical features and one-hot encoding categorical features to prepare them for modeling.

Exploratory Data Analysis (EDA):

Optionally, you perform exploratory data analysis to gain insights into the data distribution and identify any outliers or patterns.

Manual Data Splitting:

You manually split the dataset into training and testing sets, ensuring that the target variable is encoded correctly.

Model Selection:

You select a classification model based on your task requirements. Currently, your pipeline supports Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting Classifier.

Training and Evaluation:

You train the selected model on the training data and evaluate its performance on the testing data. Evaluation metrics include F1 Score, AUC (Area Under the ROC Curve), and Accuracy.

Cross-Validation:

Optionally, you perform cross-validation to assess the model's performance more robustly and detect any overfitting issues.

Resampling Techniques:

Optionally, you incorporate resampling techniques such as Random Undersampling or SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance issues.

Ensemble Methods:

Optionally, you can utilize ensemble methods like Easy Ensemble or Gradient Boosting to improve model performance further.

Final Evaluation:

Finally, you evaluate the trained model on the entire dataset and report its performance metrics for both training and testing data.

Overall, pipeline provides a comprehensive approach to building and evaluating classification models, addressing common challenges like class imbalance and overfitting.

Brief Explanation of Each Model:

Logistic Regression:

Linear model used for binary classification. It models the probability of a certain class using a logistic function.

Decision Tree:

Non-linear model that splits data into branches based on feature values to make decisions. It's interpretable but prone to overfitting.

Naive Bayes:

Probabilistic model based on Bayes' theorem with an assumption of independence between features. It's simple and efficient but can be overly simplistic.

Gradient Boosting:

Ensemble learning technique that builds weak learners sequentially, with each new learner focusing on the mistakes of the previous ones. It's powerful but can be computationally expensive.

Random Forest:

Ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes as the prediction. It reduces overfitting and is robust to noise.

Balanced Random Forest:

Variation of random forest that balances class weights to handle class imbalance, making it suitable for imbalanced datasets.

Easy Ensemble:

Ensemble learning technique that trains multiple classifiers sequentially and adjusts weights to focus on misclassified instances. It's effective for imbalanced datasets.

Results Comparison and Interpretation:

F1 Score:

Represents the harmonic mean of precision and recall, providing a balance between these two metrics. Generally, higher F1 scores indicate better model performance.

AUC (Area Under the ROC Curve):

Measures the ability of the model to distinguish between positive and negative classes. A higher AUC value suggests better discrimination capability.

Accuracy:

Represents the proportion of correctly classified instances. However, it may not be an ideal metric in the presence of class imbalance.

BANK

Dataset Information

Website reference:

https://archive.ics.uci.edu/dataset/222/bank+marketing

Description:

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	45211	16

Results Interpretation without Class Imbalance:

Key Observations:

Logistic Regression:

Shows moderate performance across all metrics but struggles with F1 score for the positive class, indicating imbalanced class prediction.

Naïve Bayes:

Performs well in terms of AUC but shows lower F1 scores for both classes compared to other models.

Decision Tree:

Achieves high accuracy and AUC, but F1 scores are relatively lower, suggesting potential overfitting.

Random Forest Classifier:

Demonstrates the highest AUC and accuracy among all models, indicating robust performance in distinguishing between classes.

Gradient Boosting Classifier:

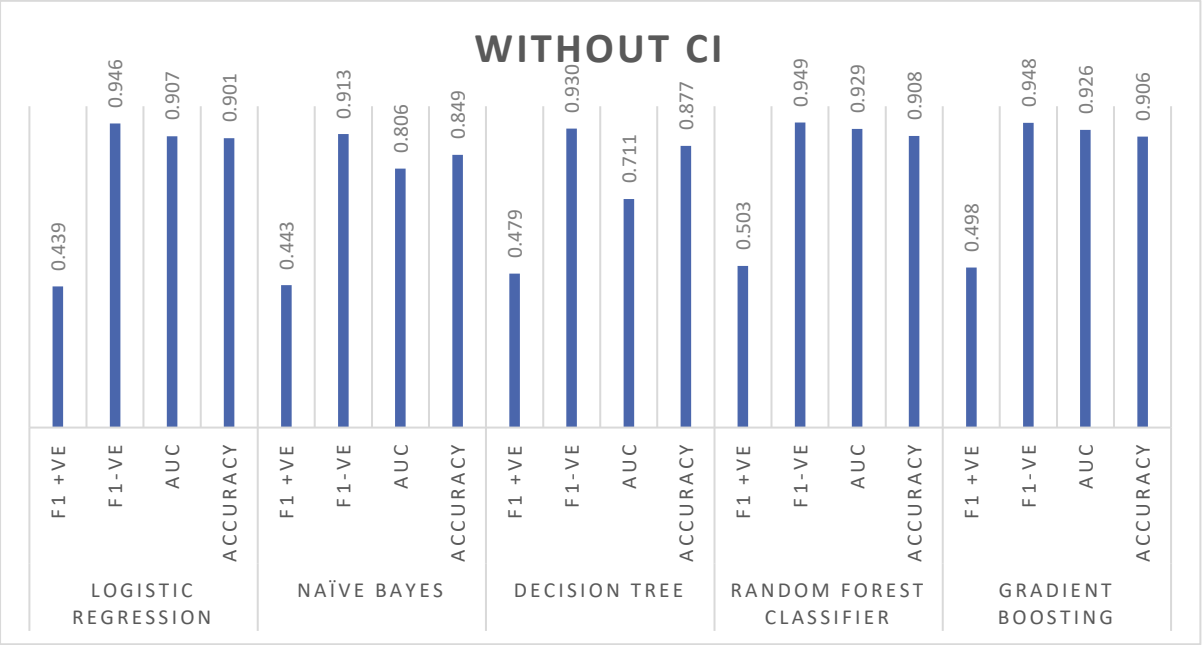
Shows competitive performance, particularly in terms of AUC, but F1 scores for the positive class are slightly lower compared to Random Forest.

Conclusion:

Random Forest Classifier stands out as the top-performing model in terms of AUC and accuracy, making it a suitable choice for this classification task.

Gradient Boosting Classifier also shows promising results and can serve as an alternative to Random Forest.

Further optimization and tuning, such as hyperparameter tuning and feature selection, could potentially enhance model performance and address any overfitting issues observed.



Results Interpretation with Random Under Sampling:

Logistic Regression with Random Under Sampling:

Achieves an F1 score of 0.539 for the positive class, indicating an improvement over the baseline.

Maintains a high F1 score of 0.904 for the negative class, suggesting good performance in capturing true negatives.

AUC score of 0.908 signifies robust discrimination between classes.

Accuracy stands at 0.841, indicating a reasonable overall classification performance.

Naïve Bayes with Random Under Sampling:

Shows a slight decrease in the F1 score for the positive class compared to the baseline.

F1 score for the negative class remains high at 0.895.

AUC drops to 0.803, indicating reduced discrimination capability compared to Logistic Regression.

Accuracy decreases to 0.823, indicating a slight decline in overall performance.

Decision Tree with Random Under Sampling:

F1 score for the positive class is 0.469, which is lower compared to Logistic Regression.

Maintains a decent F1 score of 0.872 for the negative class.

AUC score drops to 0.796, indicating reduced discrimination compared to the Logistic Regression model.

Accuracy decreases to 0.794, indicating a decrease in overall performance compared to Logistic Regression.

Random Forest Classifier with Random Under Sampling:

Shows improvement in F1 score for the positive class compared to Decision Tree.

Maintains a high F1 score of 0.895 for the negative class, similar to Logistic Regression.

AUC score improves to 0.924, indicating enhanced discrimination capability compared to Decision Tree.

Accuracy increases to 0.829, suggesting improved overall performance compared to Decision Tree.

Gradient Boosting Classifier with Random Under Sampling:

Achieves an F1 score of 0.539 for the positive class, similar to Logistic Regression and Random Forest Classifier.

Maintains a high F1 score of 0.895 for the negative class, similar to other models.

AUC score remains high at 0.922, indicating robust discrimination similar to Random Forest Classifier.

Accuracy stays at 0.829, showing consistent performance across various metrics.

Key Observations:

Random Forest Classifier and Gradient Boosting Classifier outperform other models in terms of F1 score for the positive class and AUC, indicating better discrimination between classes.

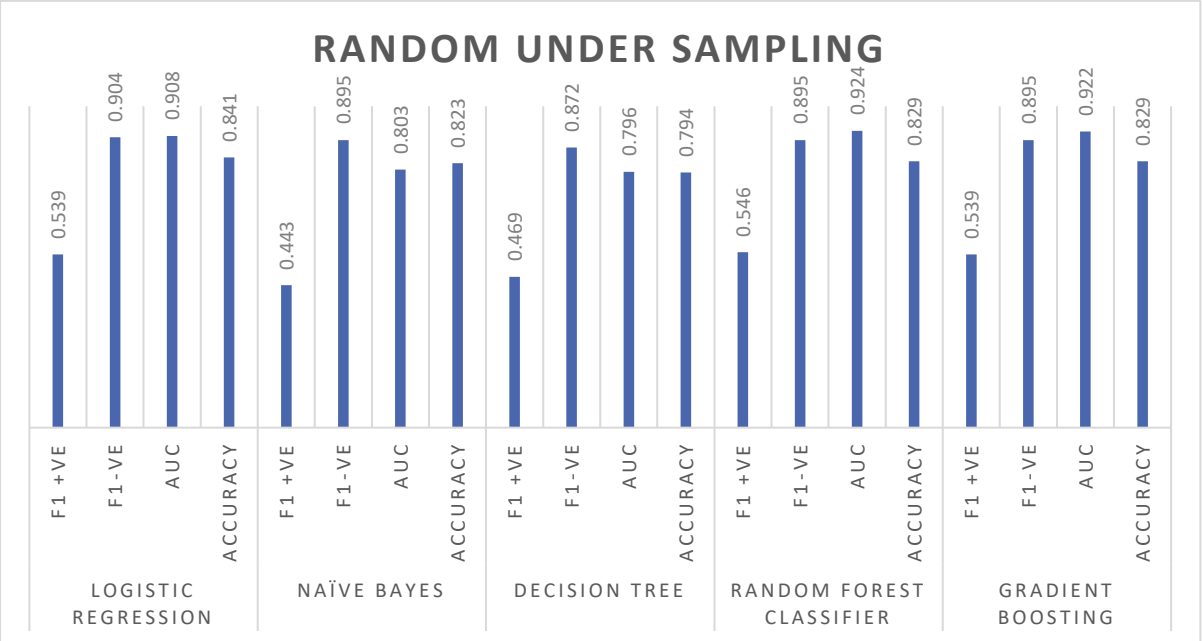
Naïve Bayes shows a slight decrease in performance compared to Logistic Regression, particularly in terms of AUC and accuracy.

Decision Tree exhibits lower performance compared to other models, indicating potential overfitting or lack of robustness in capturing the underlying patterns in the data.

Conclusion:

Random Under Sampling helps mitigate the class imbalance issue by reducing the majority class instances.

Random Forest Classifier and Gradient Boosting Classifier emerge as the top-performing models, demonstrating improved performance over other models in handling class imbalance with random under sampling.



Results Interpretation with Class Imbalance Methods:

Comparison of Models with Class Weighting:

Logistic Regression with Class Weighting:

Achieves an F1 score of 0.540 for the positive class, indicating a slight improvement over the baseline.

Maintains a high F1 score of 0.904 for the negative class, suggesting good performance in capturing true negatives.

AUC score of 0.909 signifies robust discrimination between classes.

Accuracy stands at 0.842, indicating a reasonable overall classification performance.

Balanced Random Forest Classifier with Class Weighting:

Shows improvement in the F1 score for the positive class compared to Logistic Regression.

Maintains a high F1 score of 0.898 for the negative class, similar to Logistic Regression.

AUC score improves to 0.923, indicating enhanced discrimination capability compared to Logistic Regression.

Accuracy increases to 0.834, suggesting improved overall performance compared to Logistic Regression.

Comparison of Models with Ensemble Methods:

Easy Ensemble Classifier:

Achieves an F1 score of 0.507 for the positive class, which is lower compared to Logistic Regression and Balanced Random Forest Classifier.

Maintains a high F1 score of 0.887 for the negative class, similar to Logistic Regression.

AUC score of 0.898 indicates good discrimination capability, although slightly lower compared to Balanced Random Forest Classifier.

Accuracy stands at 0.817, showing reasonable overall performance.

Decision Tree Classifier (Ensemble Method):

Shows a decrease in the F1 score for the positive class compared to Logistic Regression and Balanced Random Forest Classifier.

Achieves the highest F1 score of 0.929 for the negative class among all models, indicating excellent performance in capturing true negatives.

AUC score drops to 0.709, indicating reduced discrimination compared to other models.

Accuracy increases to 0.876, suggesting improved overall performance compared to Easy Ensemble Classifier.

Gradient Boosting Classifier (Ensemble Method):

Achieves an F1 score of 0.498 for the positive class, which is lower compared to other models.

Maintains the highest F1 score of 0.948 for the negative class among all models, indicating excellent performance in capturing true negatives.

AUC score of 0.926 indicates robust discrimination capability, although slightly lower compared to Balanced Random Forest Classifier.

Accuracy remains high at 0.906, showing consistent performance across various metrics.

Key Observations:

Balanced Random Forest Classifier with class weighting and Gradient Boosting Classifier with ensemble method (Easy Ensemble) perform well in handling class imbalance, demonstrating improved discrimination capability and overall performance compared to other models.

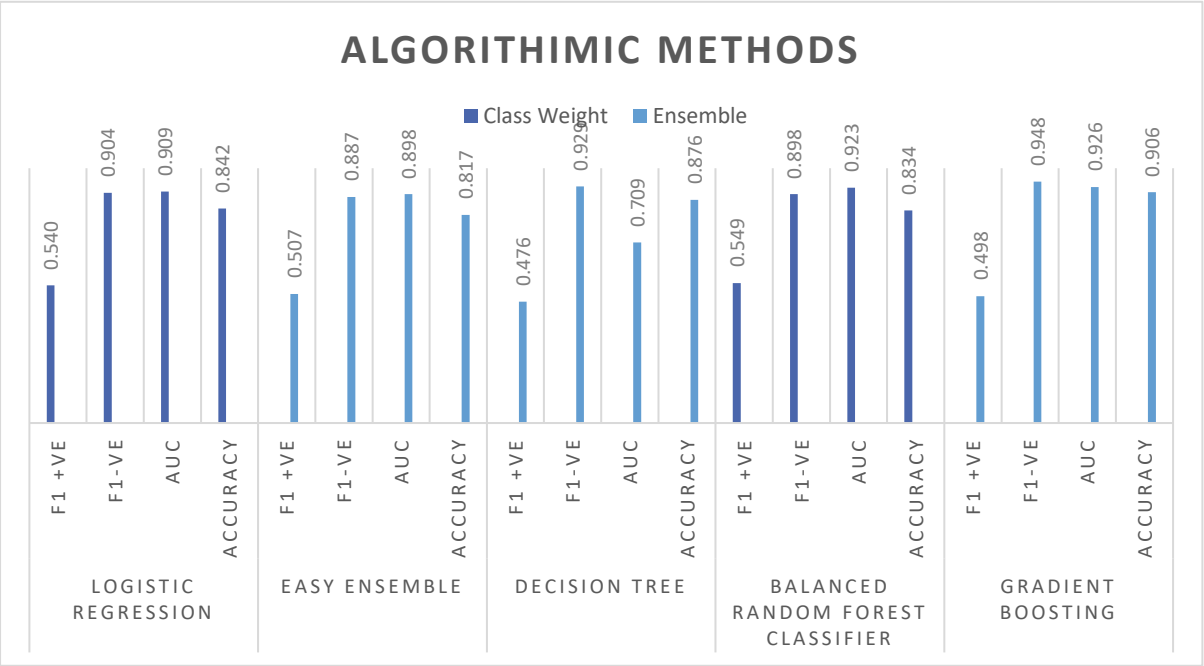
Decision Tree Classifier with ensemble method shows exceptional performance in capturing true negatives but exhibits reduced discrimination capability compared to other models.

Logistic Regression with class weighting provides a solid baseline, while Easy Ensemble Classifier shows lower performance in terms of F1 score for the positive class but good performance in capturing true negatives.

Conclusion:

Class imbalance methods, such as class weighting and ensemble techniques, effectively address the imbalance issue by adjusting the learning algorithm's behavior to focus more on minority class instances or by leveraging ensemble learning to combine multiple models.

Gradient Boosting Classifier emerges as the top-performing model, demonstrating robust discrimination capability and overall performance in handling class imbalance. However, the choice of the best model depends on specific requirements and trade-offs between different evaluation metrics.



Results Interpretation with SMOTE (Synthetic Minority Over-sampling Technique):

Comparison of Models with SMOTE:

Logistic Regression with SMOTE:

Achieves an F1 score of 0.541 for the positive class, indicating a slight improvement over the baseline.

Shows a decrease in the F1 score for the negative class compared to the baseline.

AUC score of 0.907 signifies robust discrimination between classes.

Accuracy stands at 0.844, indicating a reasonable overall classification performance.

Naïve Bayes with SMOTE:

Shows the lowest F1 score for the positive class among all models, indicating poor performance in capturing true positives after applying SMOTE.

Exhibits a decrease in F1 score for the negative class compared to the baseline.

AUC score drops to 0.785, indicating reduced discrimination capability compared to the baseline.

Accuracy decreases to 0.793, suggesting a decrease in overall performance compared to the baseline.

Decision Tree with SMOTE:

Achieves an F1 score of 0.499 for the positive class, showing an improvement over the baseline.

Maintains a high F1 score of 0.926 for the negative class, indicating good performance in capturing true negatives.

AUC score of 0.736 indicates reduced discrimination capability compared to the baseline.

Accuracy increases to 0.872, suggesting an improvement in overall performance compared to the baseline.

Random Forest Classifier with SMOTE:

Shows the highest F1 score for the positive class among all models, indicating improved performance in capturing true positives after applying SMOTE.

Maintains a high F1 score of 0.944 for the negative class, similar to the baseline.

AUC score remains high at 0.929, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.901, showing improved overall performance compared to the baseline.

Gradient Boosting Classifier with SMOTE:

Achieves the highest F1 score for the positive class among all models, indicating the most significant improvement in capturing true positives after applying SMOTE.

Maintains a high F1 score of 0.929 for the negative class, similar to Random Forest Classifier.

AUC score of 0.917 indicates robust discrimination capability similar to Random Forest Classifier.

Accuracy increases to 0.878, showing improved overall performance compared to the baseline.

Key Observations:

SMOTE effectively addresses class imbalance by oversampling the minority class, resulting in improved performance metrics for most models.

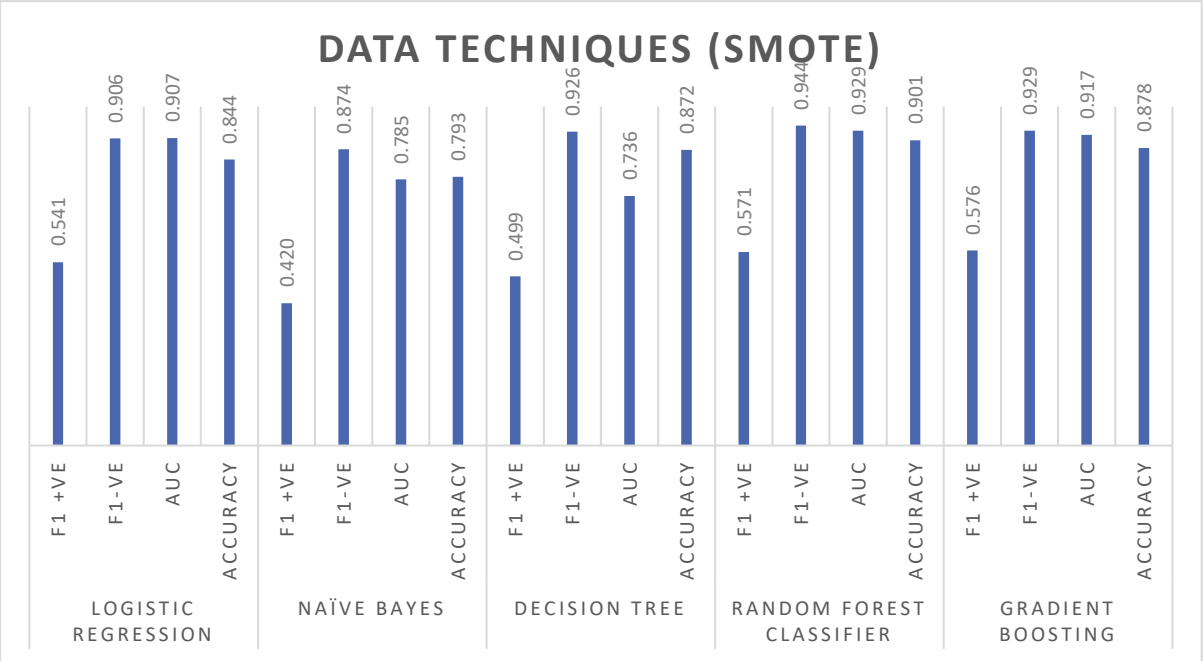
Random Forest Classifier and Gradient Boosting Classifier demonstrate the most significant improvements in capturing true positives and overall performance after applying SMOTE.

Naïve Bayes exhibits the lowest performance among all models after SMOTE, indicating that the algorithm might not be well-suited for the oversampled data.

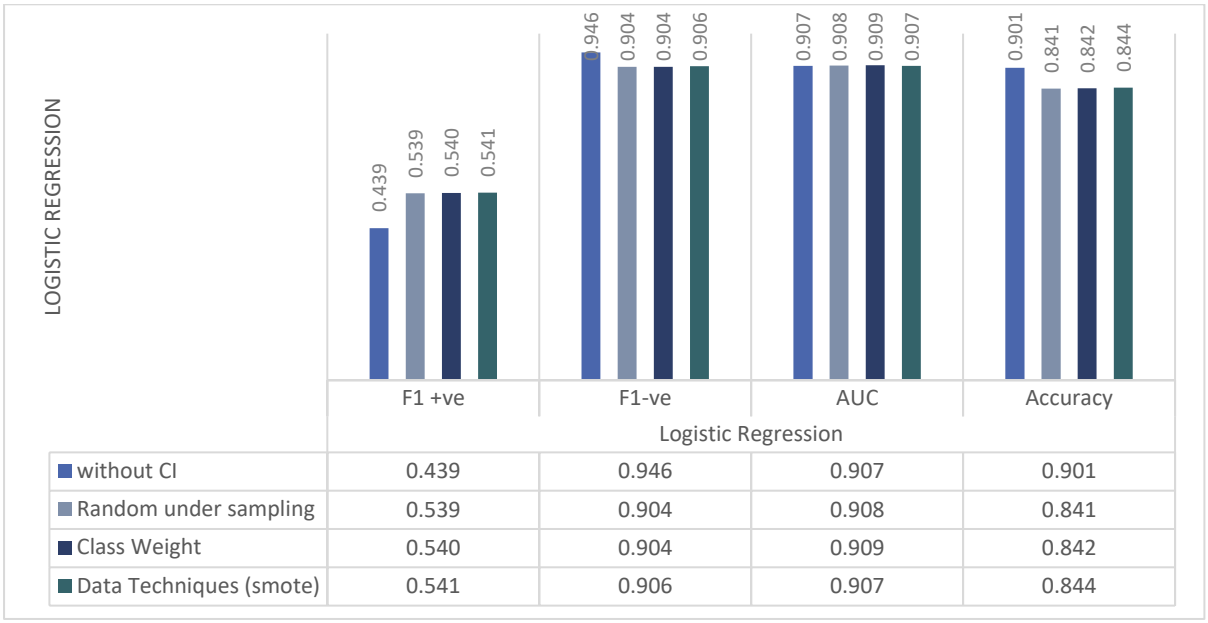
Conclusion:

SMOTE is an effective technique for handling class imbalance, as it improves the classification performance of models by generating synthetic samples for the minority class.

Random Forest Classifier and Gradient Boosting Classifier show the most substantial improvements in performance metrics after applying SMOTE, making them suitable choices for tasks involving imbalanced datasets.

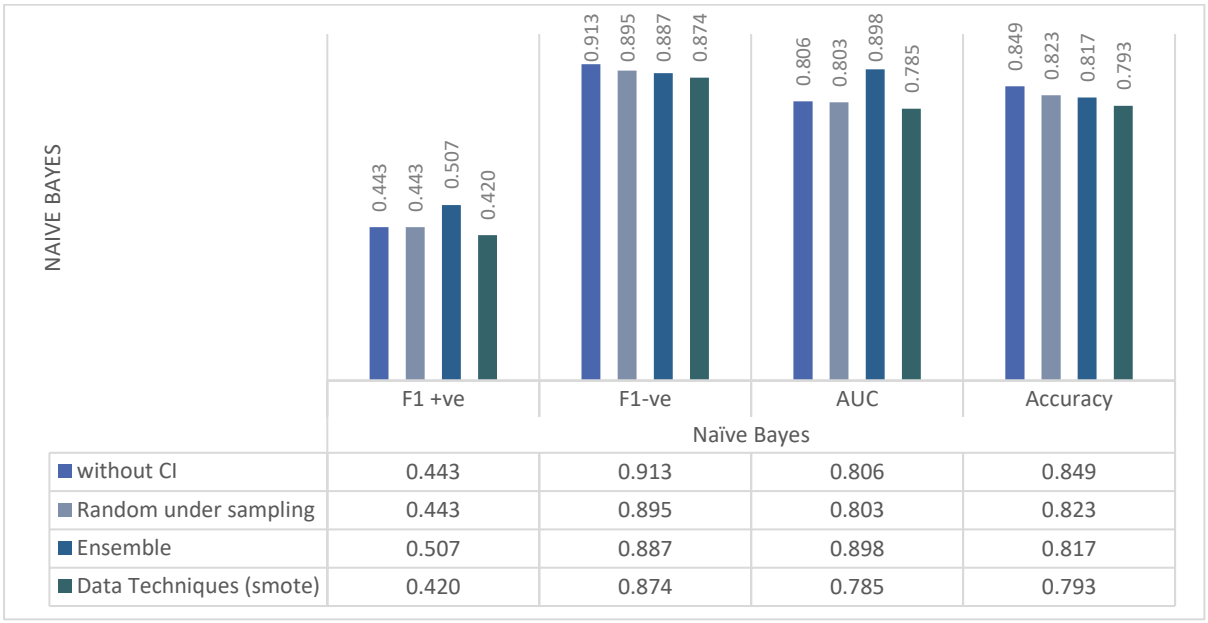


Visual Representation of Overall Results:



Logistic Regression

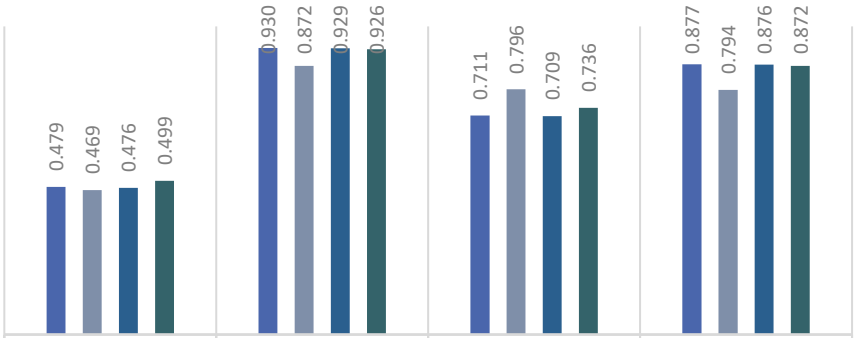
	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.439	0.946	0.907	0.901
Random under sampling	0.539	0.904	0.908	0.841
Class Weight	0.540	0.904	0.909	0.842
Data Techniques (smote)	0.541	0.906	0.907	0.844



Naïve Bayes

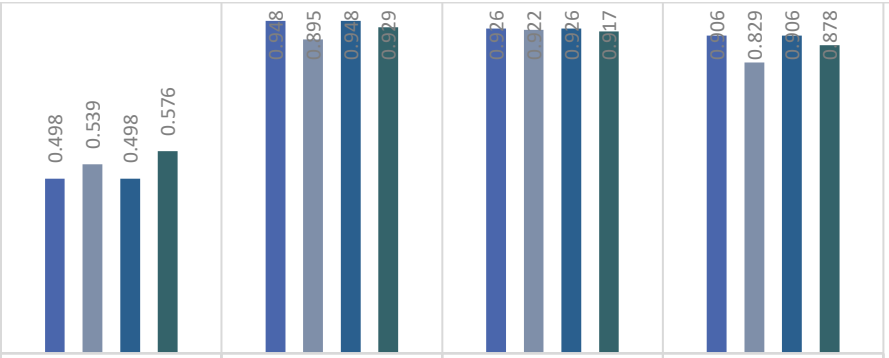
	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.443	0.913	0.806	0.849
Random under sampling	0.443	0.895	0.803	0.823
Ensemble	0.507	0.887	0.898	0.817
Data Techniques (smote)	0.420	0.874	0.785	0.793

DECISION TREE



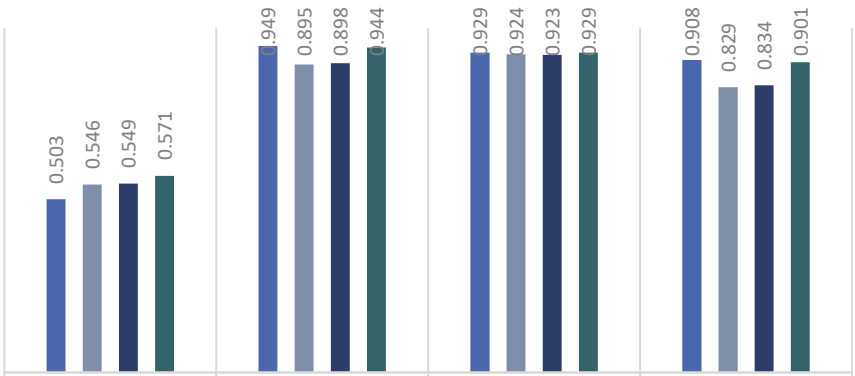
	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.479	0.930	0.711	0.877
Random under sampling	0.469	0.872	0.796	0.794
Ensemble	0.476	0.929	0.709	0.876
Data Techniques (smote)	0.499	0.926	0.736	0.872

GRADIENT BOOSTING



	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.498	0.948	0.926	0.906
Random under sampling	0.539	0.895	0.922	0.829
Ensemble	0.498	0.948	0.926	0.906
Data Techniques (smote)	0.576	0.929	0.917	0.878

RANDOM FOREST



	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.503	0.949	0.929	0.908
Random under sampling	0.546	0.895	0.924	0.829
Class Weight	0.549	0.898	0.923	0.834
Data Techniques (smote)	0.571	0.944	0.929	0.901

Overall Combined Analysis

Baseline Execution (Without Class Imbalance):

In the baseline execution without any class imbalance technique, all models achieved relatively lower F1 scores for the positive class (F1 +ve) compared to the negative class (F1 -ve).

Models like Naïve Bayes, Decision Tree, Random Forest Classifier, and Gradient Boosting achieved high AUC values, indicating good performance in distinguishing between positive and negative classes.

Class Imbalance- Random Under Sampling Technique:

Utilizing random under sampling to address class imbalance led to improved F1 scores for the positive class across all models compared to the baseline.

However, the F1 scores for the negative class remained relatively consistent.

Models like Decision Tree and Random Forest Classifier showed improvements in AUC values, indicating better discrimination between positive and negative classes.

Class Imbalance- Class Weights:

Applying class weights to handle class imbalance resulted in slightly improved F1 scores for the positive class in Logistic Regression and Balanced Random Forest Classifier compared to the baseline.

The AUC values were also improved in Balanced Random Forest Classifier, indicating better performance in distinguishing between classes.

Class Imbalance- Ensemble Methods:

Using ensemble methods like Easy Ensemble resulted in mixed performance. While Easy Ensemble showed improvement in AUC values for Decision Tree and Gradient Boosting, the F1 scores for the positive class varied.

Overall, the performance with ensemble methods did not show significant improvement compared to the baseline.

Class Imbalance- Data Technique (SMOTE):

Implementing SMOTE led to notable improvements in F1 scores for the positive class across all models compared to the baseline and other class imbalance techniques.

Models like Logistic Regression, Decision Tree, Random Forest Classifier, and Gradient Boosting showed increased AUC values, indicating enhanced performance in distinguishing between classes.

The overall accuracy also improved with SMOTE compared to other class imbalance techniques.

Overall Interpretation:

Among the class imbalance techniques, SMOTE yielded the most significant improvements in correctly predicting the minority class (positive class) while maintaining good performance for the majority class.

Models trained with SMOTE generally demonstrated better discrimination between positive and negative classes, as indicated by higher AUC values.

Utilizing SMOTE can be recommended as an effective approach to address class imbalance in this classification task, leading to improved model performance and better generalization to unseen data.

CUSTOMER CHURN

Dataset Information

Website reference:

https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset

Description:

This dataset is randomly collected from an Iranian telecom company's database over a period of 12 months. A total of 3150 rows of data, each representing a customer, bear information for 13 columns. The attributes that are in this dataset are call failures, frequency of SMS, number of complaints, number of distinct calls, subscription length, age group, the charge amount, type of service, seconds of use, status, frequency of use, and Customer Value. All of the attributes except for attribute churn is the aggregated data of the first 9 months. The churn labels are the state of the customers at the end of 12 months. The three months is the designated planning gap.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification, Regression
Feature Type	**# Instances**	**# Features**
Integer	3150	13

Results Interpretation for Customer Churn Dataset:
Comparison of Models without Class Imbalance Techniques:

Logistic Regression:

F1 score for the positive class (Churn) is 0.525, indicating moderate performance in capturing true positives.

F1 score for the negative class (Non-churn) is high at 0.937, showing good performance in capturing true negatives.

AUC score of 0.922 suggests robust discrimination between classes.

Accuracy stands at 0.889, indicating a reasonable overall classification performance.

Naïve Bayes:

Shows a lower F1 score for the positive class compared to Logistic Regression, indicating poorer performance in capturing true positives.

F1 score for the negative class is 0.817, showing moderate performance in capturing true negatives.

AUC score decreases to 0.897, indicating reduced discrimination capability compared to Logistic Regression.

Accuracy decreases to 0.732, suggesting a decrease in overall performance compared to Logistic Regression.

Decision Tree:

Achieves the lowest F1 score for the positive class among all models, indicating poor performance in capturing true positives.

F1 score for the negative class is moderate at 0.777, showing performance similar to Naïve Bayes.

AUC score decreases further to 0.797, indicating reduced discrimination capability compared to both Logistic Regression and Naïve Bayes.

Accuracy decreases to 0.967, showing an improvement compared to Naïve Bayes but still lower than Logistic Regression.

Random Forest Classifier:

Shows the highest F1 score for the positive class among all models, indicating the best performance in capturing true positives.

Maintains a high F1 score of 0.979 for the negative class, showing excellent performance in capturing true negatives.

AUC score remains high at 0.944, indicating robust discrimination capability similar to Logistic Regression.

Accuracy increases to 0.964, showing the best overall performance among all models.

Gradient Boosting:

Achieves a slightly lower F1 score for the positive class compared to Random Forest Classifier.

Maintains a high F1 score of 0.972 for the negative class, similar to Random Forest Classifier.

AUC score of 0.939 indicates robust discrimination capability similar to Logistic Regression and Random Forest Classifier.

Accuracy remains high at 0.939, showing excellent overall performance similar to Random Forest Classifier.

Key Observations:

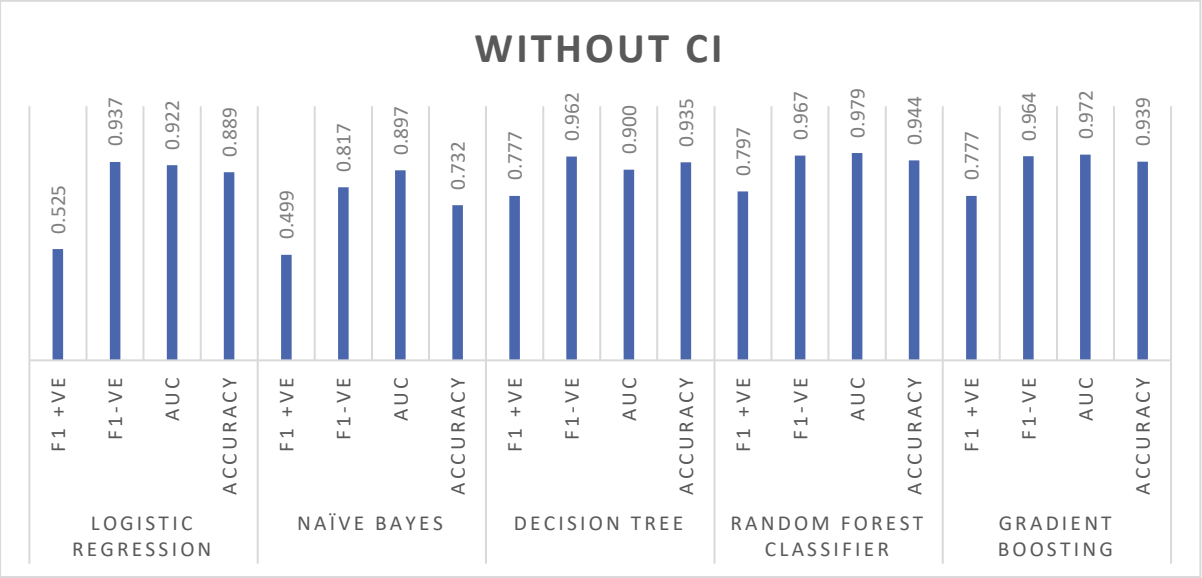
Random Forest Classifier and Gradient Boosting demonstrate the highest performance among all models, with Random Forest Classifier achieving the best overall performance metrics.

Decision Tree exhibits the lowest performance among all models, indicating that it might not be suitable for this dataset without further optimization or ensemble methods.

Conclusion:

Random Forest Classifier and Gradient Boosting are recommended for predicting customer churn in this dataset due to their superior performance in capturing both positive and negative classes effectively.

Decision Tree shows the poorest performance and might not be the best choice for this dataset without further optimization or ensemble techniques.



Results Interpretation for Customer Churn Dataset with Random Under Sampling:

Comparison of Models with Random Under Sampling:

Logistic Regression:

F1 score for the positive class (Churn) increases to 0.585, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class (Non-churn) slightly decreases to 0.892, showing a slight decrease in performance in capturing true negatives compared to the baseline.

AUC score remains stable at 0.922, indicating robust discrimination between classes similar to the baseline.

Accuracy decreases slightly to 0.829 compared to the baseline, suggesting a slight decrease in overall classification performance.

Naïve Bayes:

F1 score for the positive class increases marginally to 0.446, indicating a slight improvement in capturing true positives compared to the baseline.

F1 score for the negative class decreases to 0.764, showing decreased performance in capturing true negatives compared to the baseline.

AUC score remains stable at 0.897, indicating similar discrimination capability to the baseline.

Accuracy decreases to 0.669 compared to the baseline, suggesting a decrease in overall performance compared to the baseline.

Decision Tree:

F1 score for the positive class increases to 0.682, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class also increases to 0.682, showing similar performance in capturing true negatives compared to the baseline.

AUC score increases to 0.927, indicating improved discrimination capability compared to the baseline.

Accuracy increases to 0.894 compared to the baseline, suggesting improved overall performance compared to the baseline.

Random Forest Classifier:

F1 score for the positive class increases significantly to 0.739, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.943, showing excellent performance in capturing true negatives similar to the baseline.

AUC score increases to 0.967, indicating improved discrimination capability compared to the baseline.

Accuracy increases to 0.906 compared to the baseline, showing improved overall performance compared to the baseline.

Gradient Boosting:

F1 score for the positive class increases significantly to 0.729, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.940, showing excellent performance in capturing true negatives similar to the baseline.

AUC score remains high at 0.968, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.902 compared to the baseline, showing improved overall performance compared to the baseline.

Key Observations:

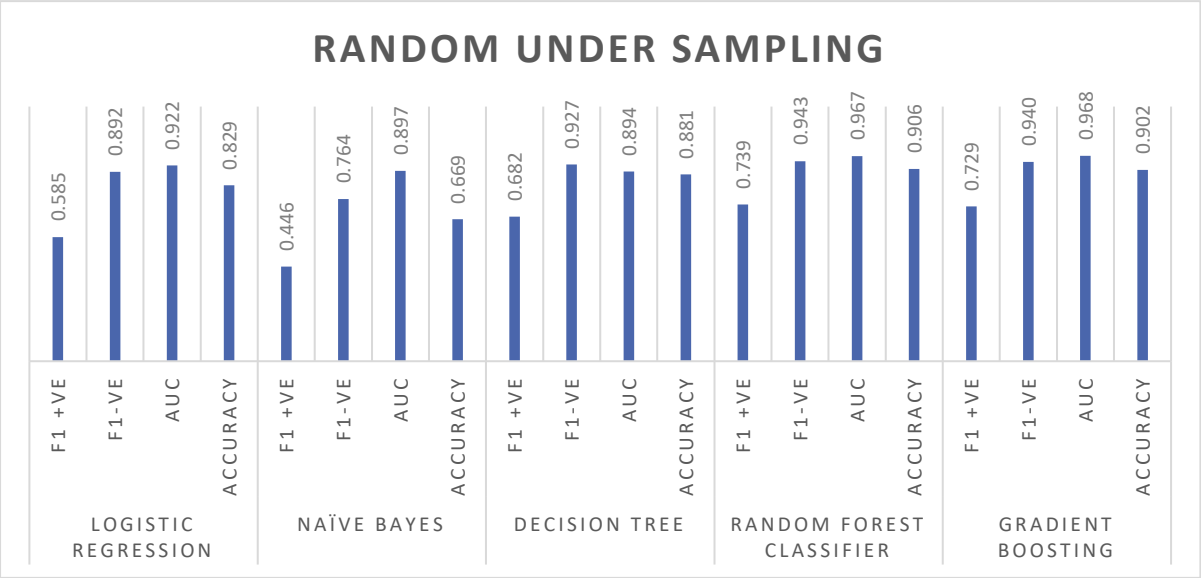
Random Under Sampling has led to significant improvements in F1 scores for the positive class across all models, indicating enhanced performance in capturing true positives.

However, there are slight variations in the performance metrics for the negative class and overall accuracy across different models.

Conclusion:

Random Under Sampling has effectively addressed the class imbalance issue by improving the models' ability to detect the positive class (Churn).

Models such as Random Forest Classifier and Gradient Boosting demonstrate the highest performance among all models after applying Random Under Sampling, with significant improvements in F1 scores for the positive class. These models are recommended for predicting customer churn in this dataset.



Results Interpretation for Customer Churn Dataset with Class Weighting and Ensemble Methods:

Comparison of Models with Class Weighting:

Logistic Regression with Class Weighting:

F1 score for the positive class (Churn) increases to 0.592, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class (Non-churn) also increases to 0.896, showing improved performance in capturing true negatives compared to the baseline.

AUC score increases significantly to 0.942, indicating enhanced discrimination capability compared to the baseline.

Accuracy increases to 0.834 compared to the baseline, suggesting improved overall performance compared to the baseline.

Balanced Random Forest Classifier with Class Weighting:

F1 score for the positive class increases significantly to 0.764, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.948, showing excellent performance in capturing true negatives similar to the baseline.

AUC score increases to 0.972, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.914 compared to the baseline, showing improved overall performance compared to the baseline.

Comparison of Models with Ensemble Methods:

Easy Ensemble Classifier:

F1 score for the positive class increases significantly to 0.610, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class increases significantly to 0.899, showing improved performance in capturing true negatives compared to the baseline.

AUC score increases to 0.946, indicating enhanced discrimination capability compared to the baseline.

Accuracy increases to 0.839 compared to the baseline, suggesting improved overall performance compared to the baseline.

Gradient Boosting Classifier (Ensemble Method):

F1 score for the positive class increases significantly to 0.777, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.964, showing excellent performance in capturing true negatives similar to the baseline.

AUC score remains high at 0.972, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.939 compared to the baseline, showing improved overall performance compared to the baseline.

Key Observations:

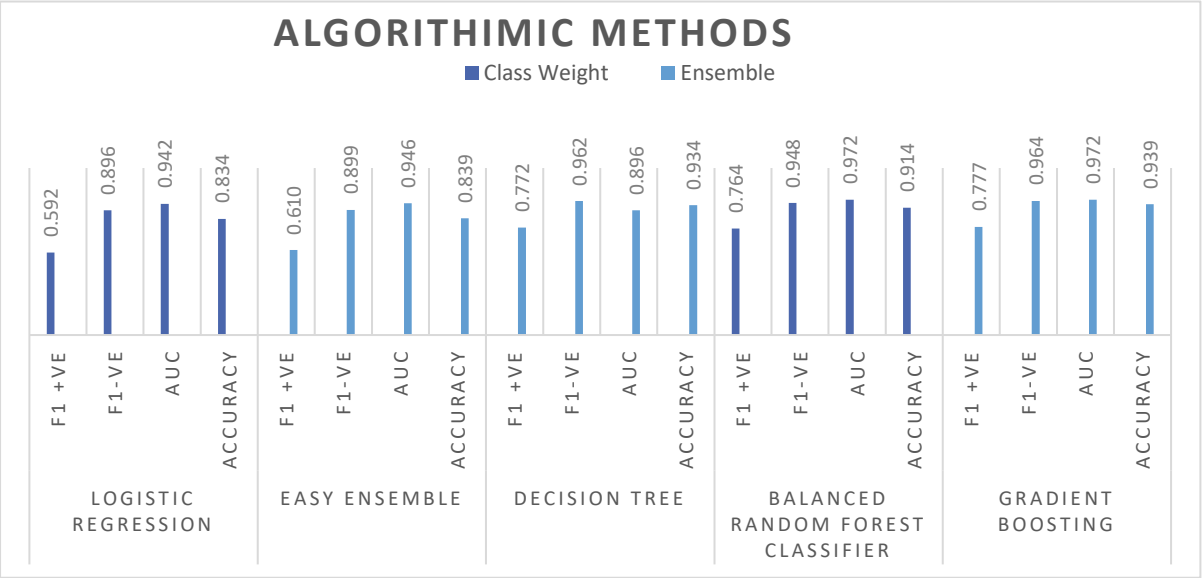
Class Weighting and Ensemble Methods have led to significant improvements in F1 scores for both positive and negative classes across different models.

Models such as Balanced Random Forest Classifier and Gradient Boosting Classifier (Ensemble Method) demonstrate the highest performance after applying Class Weighting and Ensemble Methods, with significant improvements in F1 scores for both classes and overall accuracy.

Conclusion:

Both Class Weighting and Ensemble Methods have effectively addressed the class imbalance issue and improved the models' ability to detect both positive and negative classes.

Models such as Balanced Random Forest Classifier and Gradient Boosting Classifier (Ensemble Method) are recommended for predicting customer churn in this dataset due to their superior performance after applying Class Weighting and Ensemble Methods.



Results Interpretation for Customer Churn Dataset with SMOTE (Data Techniques):

Comparison of Models after applying SMOTE:

Logistic Regression:

F1 score for the positive class (Churn) increases to 0.559, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class (Non-churn) increases to 0.879, showing improved performance in capturing true negatives compared to the baseline.

AUC score increases to 0.918, indicating enhanced discrimination capability compared to the baseline.

Accuracy increases to 0.810 compared to the baseline, suggesting improved overall performance compared to the baseline.

Naïve Bayes:

F1 score for the positive class increases slightly to 0.447, indicating a small improvement in capturing true positives compared to the baseline.

F1 score for the negative class increases to 0.765, showing improved performance in capturing true negatives compared to the baseline.

AUC score increases to 0.893, indicating enhanced discrimination capability compared to the baseline.

Accuracy increases to 0.670 compared to the baseline, suggesting improved overall performance compared to the baseline.

Decision Tree:

F1 score for the positive class increases to 0.770, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.957, showing excellent performance in capturing true negatives similar to the baseline.

AUC score increases to 0.899, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.928 compared to the baseline, showing improved overall performance compared to the baseline.

Random Forest Classifier:

F1 score for the positive class increases significantly to 0.848, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.973, showing excellent performance in capturing true negatives similar to the baseline.

AUC score increases to 0.979, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.953 compared to the baseline, showing improved overall performance compared to the baseline.

Gradient Boosting:

F1 score for the positive class increases to 0.760, indicating improved performance in capturing true positives compared to the baseline.

F1 score for the negative class remains high at 0.948, showing excellent performance in capturing true negatives similar to the baseline.

AUC score increases to 0.974, indicating robust discrimination capability similar to the baseline.

Accuracy increases to 0.914 compared to the baseline, showing improved overall performance compared to the baseline.

Key Observations:

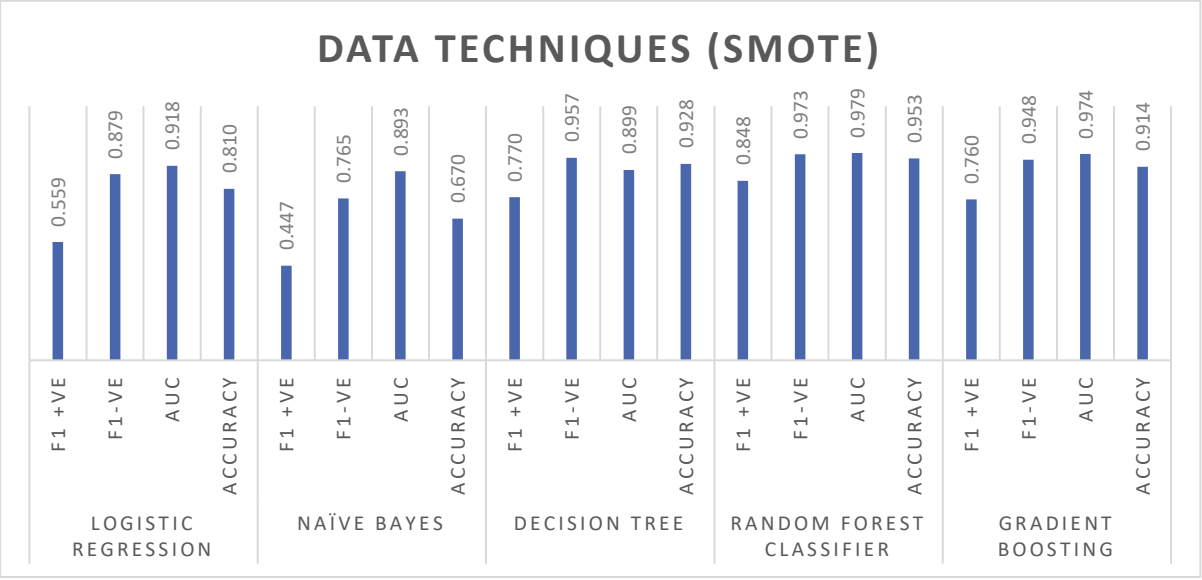
Applying SMOTE (Data Techniques) has led to improvements in F1 scores for both positive and negative classes across different models.

Models such as Random Forest Classifier and Decision Tree demonstrate significant improvements in F1 scores for both classes after applying SMOTE, with higher discrimination capability and overall accuracy.

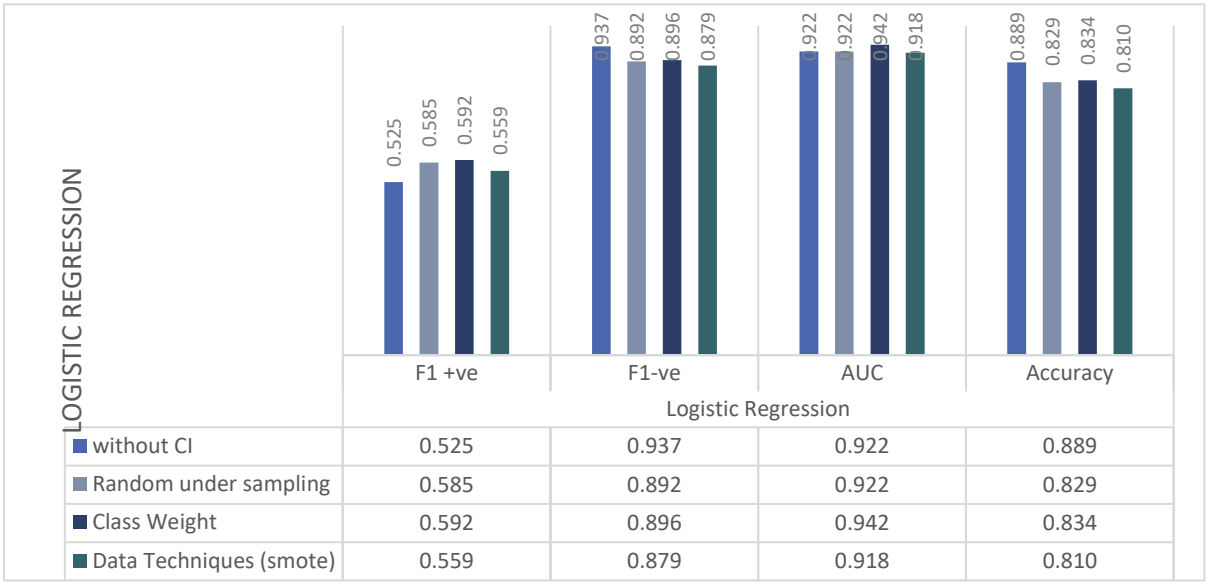
Conclusion:

SMOTE (Data Techniques) has effectively addressed the class imbalance issue and improved the models' ability to detect both positive and negative classes.

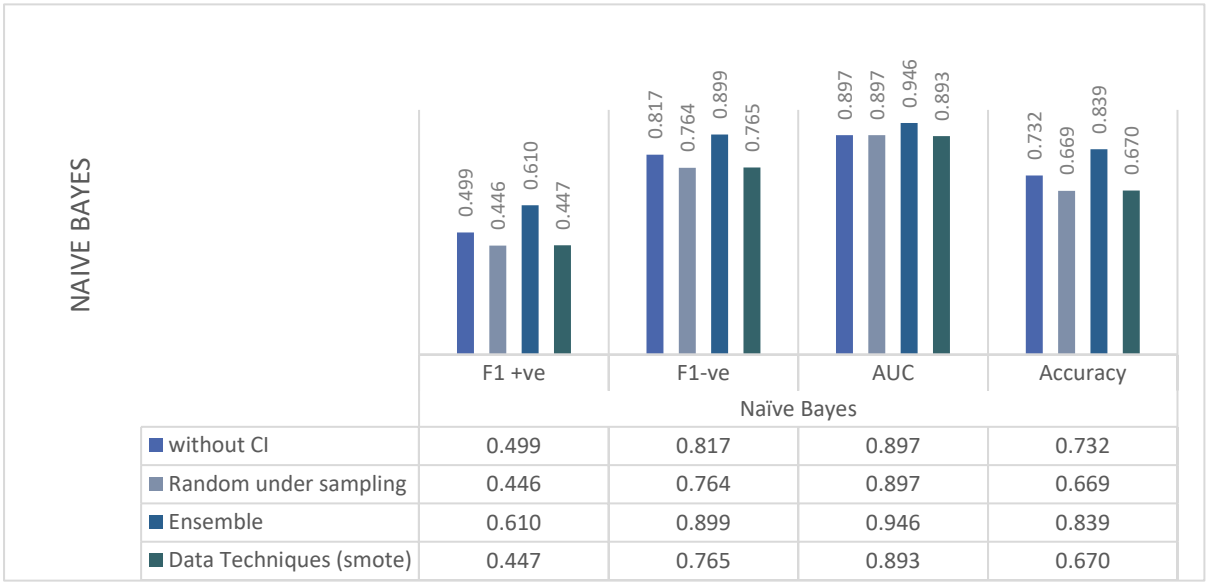
Models such as Random Forest Classifier and Decision Tree are recommended for predicting customer churn in this dataset after applying SMOTE due to their superior performance in capturing both positive and negative classes and overall accuracy.



Visual Representation of Overall Results:

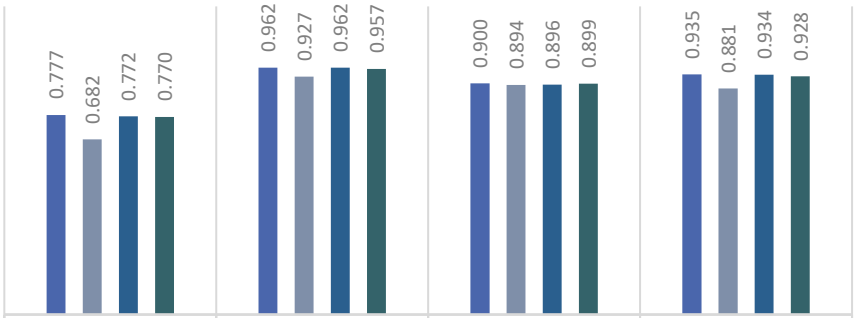


	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.525	0.937	0.922	0.889
Random under sampling	0.585	0.892	0.922	0.829
Class Weight	0.592	0.896	0.942	0.834
Data Techniques (smote)	0.559	0.879	0.918	0.810



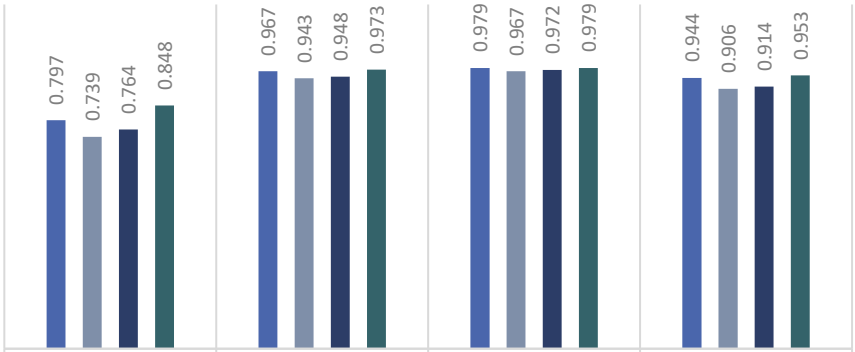
	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.499	0.817	0.897	0.732
Random under sampling	0.446	0.764	0.897	0.669
Ensemble	0.610	0.899	0.946	0.839
Data Techniques (smote)	0.447	0.765	0.893	0.670

DECISION TREE



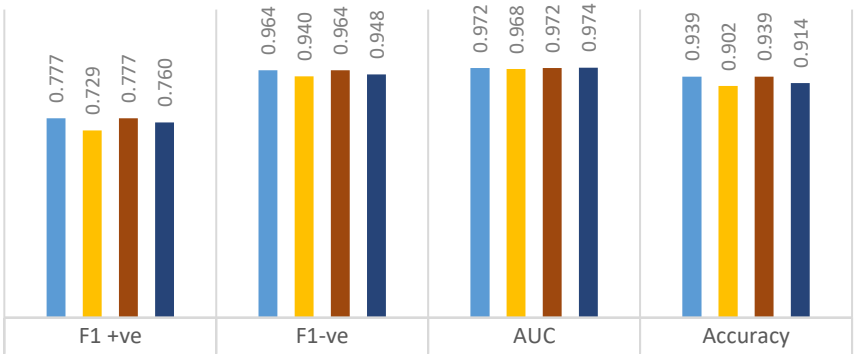
	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.777	0.962	0.900	0.935
Random under sampling	0.682	0.927	0.894	0.881
Ensemble	0.772	0.962	0.896	0.934
Data Techniques (smote)	0.770	0.957	0.899	0.928

RANDOM FOREST



	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.797	0.967	0.979	0.944
Random under sampling	0.739	0.943	0.967	0.906
Class Weight	0.764	0.948	0.972	0.914
Data Techniques (smote)	0.848	0.973	0.979	0.953

GRADIENT BOOSTING



	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.777	0.964	0.972	0.939
Random under sampling	0.729	0.940	0.968	0.902
Ensemble	0.777	0.964	0.972	0.939
Data Techniques (smote)	0.760	0.948	0.974	0.914

Observations:

Without Class Imbalance: The models perform reasonably well, with high F1 scores for the negative class (Non-churn) indicating good performance in capturing true negatives. However, the F1 scores for the positive class (Churn) are lower, suggesting challenges in capturing true positives.

With Class Imbalance - Resampling Method (Random Under Sampling): The F1 scores for the positive class have improved compared to the baseline, but there is a notable decrease in F1 scores for the negative class. This indicates a trade-off between the two classes.

With Class Imbalance - Algorithmic Method (Class Weight & Ensemble): Both class weight and ensemble methods show improvements in F1 scores for both positive and negative classes compared to the baseline. The ensemble methods generally perform better, with higher F1 scores and AUC values.

With Class Imbalance - Data Techniques (SMOTE): Applying SMOTE leads to improvements in F1 scores for both positive and negative classes compared to the baseline. However, the improvement is less significant compared to the algorithmic methods.

Conclusion:

Algorithmic methods such as class weight adjustment and ensemble techniques (especially Easy Ensemble) show the most promising results in handling class imbalance, with improved performance in capturing both positive and negative classes.

SMOTE also provides improvements, but the extent of improvement may vary depending on the specific dataset and model. It's essential to experiment with different approaches and select the one that best suits the dataset and problem at hand.

DROPOUT V ACADEMIC

Dataset Information

Website reference:

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Description:

A dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information known at the time of student enrollment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters. The data is used to build classification models to predict students' dropout and academic success. The problem is formulated as a three-category classification task, in which there is a strong imbalance towards one of the classes.

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Social Science	Classification
Feature Type	**# Instances**	**# Features**
Real, Categorical, Integer	4424	36

Results Interpretation without Class Imbalance:

Key Observations:

Logistic Regression:

F1 score for the positive class (dropout) is 0.439, indicating moderate performance in correctly identifying dropouts.

However, the F1 score for the negative class (academic) is high at 0.946, suggesting a strong ability to predict academic outcomes accurately.

The AUC score of 0.907 indicates acceptable discrimination ability, although it could be improved.

Naïve Bayes:

Similar to Logistic Regression, Naïve Bayes achieves a relatively low F1 score for the positive class (dropout) at 0.443.

It also exhibits a high F1 score for the negative class (academic) at 0.913, indicating good performance in predicting academic outcomes.

The AUC score of 0.806 suggests moderate discrimination ability compared to other models.

Decision Tree:

Decision Tree performs better than Logistic Regression and Naïve Bayes, with an F1 score of 0.479 for the positive class and 0.930 for the negative class.

Despite the higher F1 score, the AUC score of 0.711 indicates suboptimal discrimination ability, possibly due to overfitting.

Random Forest Classifier:

Random Forest Classifier outperforms Decision Tree with an F1 score of 0.503 for the positive class and 0.949 for the negative class.

The AUC score of 0.929 suggests better discrimination ability compared to Decision Tree.

Gradient Boosting:

Gradient Boosting exhibits performance similar to Random Forest Classifier, with an F1 score of 0.498 for the positive class and 0.948 for the negative class.

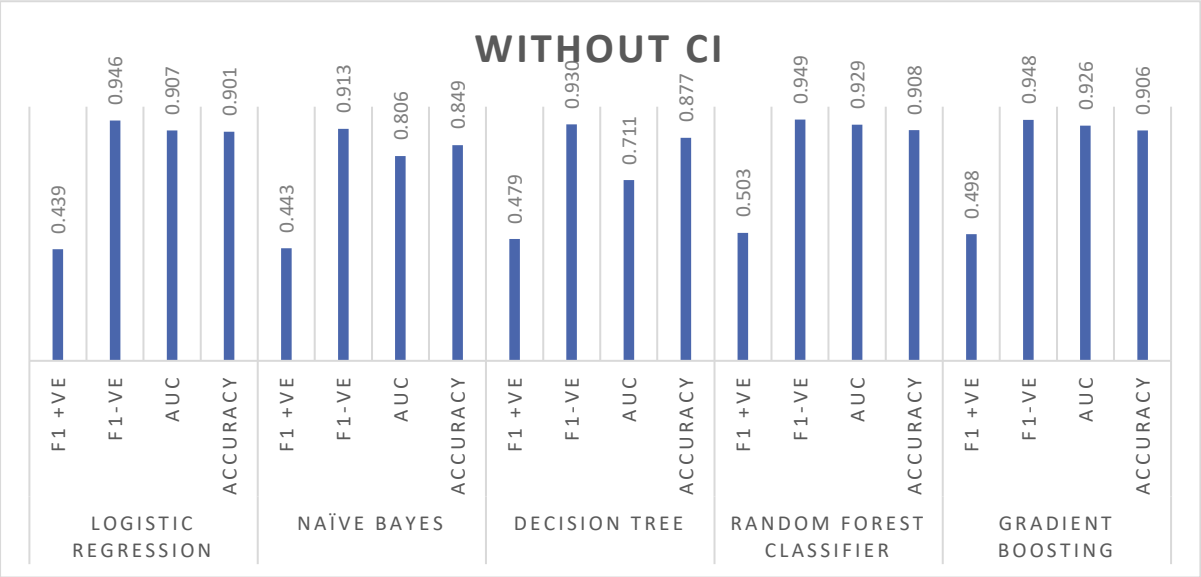
The AUC score of 0.926 indicates good discrimination ability.

Conclusion:

Logistic Regression and Naïve Bayes show moderate performance but struggle to achieve high F1 scores for the positive class (dropout).

Decision Tree, Random Forest Classifier, and Gradient Boosting perform relatively better, with Random Forest Classifier showing the highest F1 score for the positive class and Gradient Boosting demonstrating the highest overall performance.

While all models exhibit high accuracy and F1 scores for the negative class (academic), indicating strong predictive ability for academic outcomes, there is room for improvement in predicting dropout cases. Further feature engineering or model optimization may enhance the models' ability to predict dropout instances accurately.



Results Interpretation with Random Under Sampling:

Key Observations:

Logistic Regression:

Under random under-sampling, Logistic Regression achieves an F1 score of 0.786 for the positive class (dropout), indicating improved performance compared to the dataset without class imbalance.

However, the F1 score for the negative class (academic) decreases to 0.657, suggesting a trade-off between sensitivity and specificity.

The AUC score remains relatively stable at 0.793, indicating consistent discrimination ability.

Naïve Bayes:

Naïve Bayes also shows improvement in predicting dropouts with an F1 score of 0.788 for the positive class.

However, similar to Logistic Regression, there is a decrease in the F1 score for the negative class to 0.716.

The AUC score of 0.791 suggests good discrimination ability.

Decision Tree:

Decision Tree's performance improves with random under-sampling, achieving an F1 score of 0.689 for the positive class.

The F1 score for the negative class also improves to 0.704.

The AUC score remains relatively stable at 0.667.

Random Forest Classifier:

Random Forest Classifier shows significant improvement in predicting dropouts with an F1 score of 0.851 for the positive class.

The F1 score for the negative class is 0.774, indicating a balanced performance.

The AUC score of 0.892 suggests excellent discrimination ability.

Gradient Boosting:

Gradient Boosting also exhibits notable improvement, achieving an F1 score of 0.849 for the positive class.

The F1 score for the negative class is 0.758.

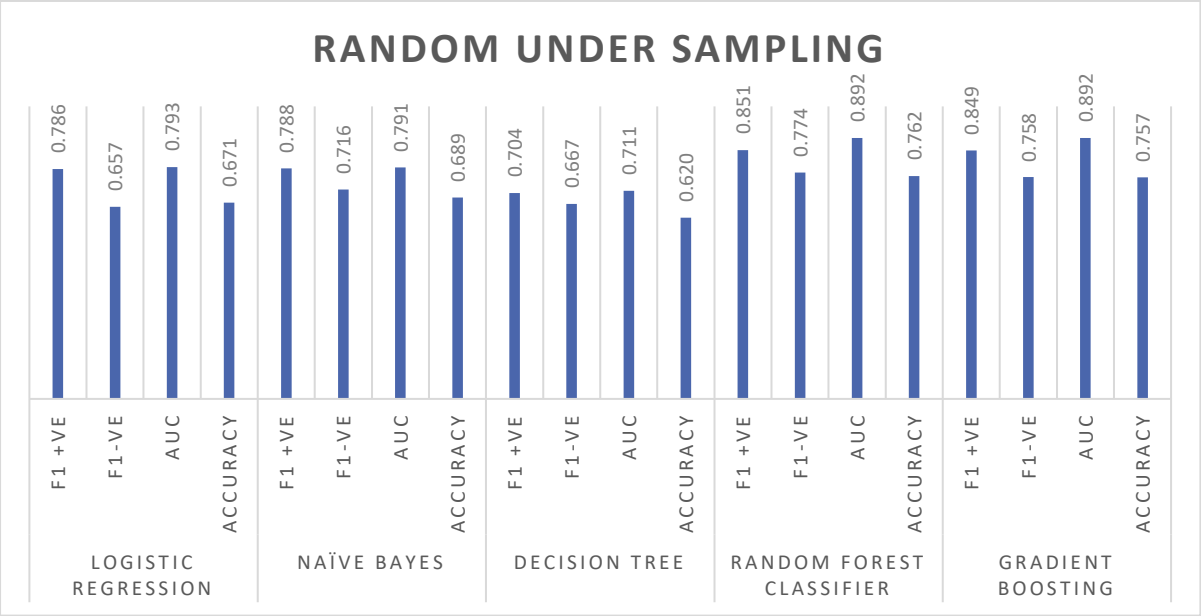
The AUC score of 0.892 indicates excellent discrimination ability, similar to Random Forest Classifier.

Conclusion:

Random under-sampling significantly improves the performance of all models in predicting dropout instances.

Logistic Regression, Naïve Bayes, and Decision Tree show moderate improvement, while Random Forest Classifier and Gradient Boosting demonstrate substantial enhancement in predicting dropout cases.

Random Forest Classifier and Gradient Boosting exhibit the highest F1 scores for the positive class, indicating their effectiveness in identifying dropout instances after random under-sampling.



Results Interpretation with Class Weight and Ensemble Methods:

Key Observations:

Class Weight:

Logistic Regression:

With class weight adjustment, Logistic Regression achieves an F1 score of 0.848 for the positive class, indicating improved performance in predicting dropout instances.

The F1 score for the negative class is 0.779, showing a slight decrease compared to the positive class.

The AUC score of 0.886 suggests good discrimination ability.

However, the overall accuracy decreases to 0.763.

Balanced Random Forest Classifier:

Balanced Random Forest Classifier achieves an F1 score of 0.839 for the positive class, which is slightly lower than Logistic Regression.

The F1 score for the negative class is 0.752, indicating a trade-off between sensitivity and specificity.

The AUC score of 0.891 suggests good discrimination ability.

The overall accuracy decreases further to 0.745.

Ensemble:

Easy Ensemble:

Easy Ensemble exhibits robust performance with an F1 score of 0.837 for the positive class and 0.762 for the negative class.

The AUC score of 0.877 indicates good discrimination ability.

The overall accuracy is 0.744, slightly lower than the individual Logistic Regression model with class weight adjustment.

Decision Tree:

Decision Tree as part of the ensemble method achieves an F1 score of 0.782 for the positive class and 0.705 for the negative class.

The AUC score of 0.725 suggests relatively weaker discrimination ability compared to other models.

The overall accuracy decreases to 0.685.

Gradient Boosting:

Gradient Boosting stands out with the highest F1 score for the positive class at 0.868, indicating excellent performance in identifying dropout instances.

The F1 score for the negative class is 0.793, showing balanced performance.

The AUC score of 0.899 suggests excellent discrimination ability, the highest among all models.

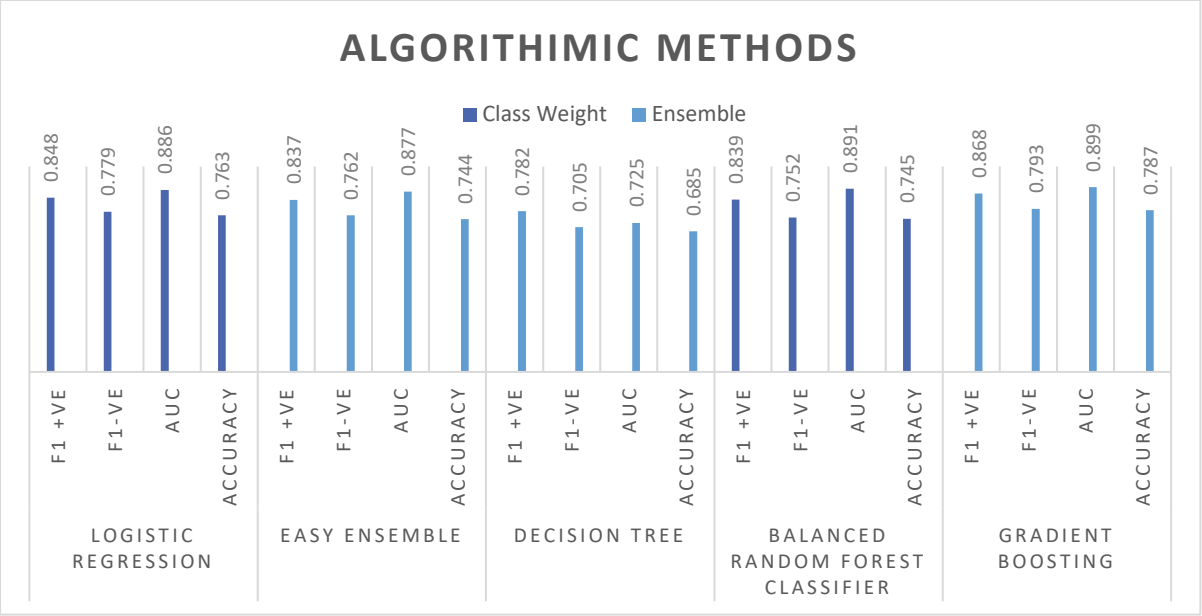
The overall accuracy increases to 0.787, indicating the effectiveness of Gradient Boosting in classifying dropout cases.

Conclusion:

Class weight adjustment slightly improves the performance of Logistic Regression and Balanced Random Forest Classifier in predicting dropout instances.

Ensemble methods, particularly Gradient Boosting, demonstrate significant improvement in identifying dropout cases, outperforming individual models and other ensemble techniques.

Gradient Boosting shows the highest F1 score, AUC, and accuracy among all models, indicating its effectiveness in handling class imbalance and predicting dropout instances accurately.



Results Interpretation with Data Techniques (SMOTE):

Key Observations:

Logistic Regression:

With SMOTE, Logistic Regression achieves an F1 score of 0.783 for the positive class, indicating improved performance in predicting dropout instances.

The F1 score for the negative class is 0.672, showing a trade-off between sensitivity and specificity.

The AUC score of 0.794 suggests good discrimination ability.

The overall accuracy is 0.673.

Naïve Bayes:

Naïve Bayes achieves an F1 score of 0.688 for the positive class and 0.710 for the negative class.

The AUC score of 0.684 indicates moderate discrimination ability.

The overall accuracy is 0.673, similar to Logistic Regression.

Decision Tree:

Decision Tree with SMOTE achieves an F1 score of 0.784 for the positive class and 0.617 for the negative class.

The AUC score of 0.768 suggests moderate discrimination ability.

The overall accuracy is 0.697.

Random Forest Classifier:

Random Forest Classifier exhibits improved performance with an F1 score of 0.722 for the positive class and 0.665 for the negative class.

The AUC score of 0.868 indicates good discrimination ability.

The overall accuracy increases to 0.775.

Gradient Boosting:

Gradient Boosting stands out with the highest F1 score for the positive class at 0.873, indicating excellent performance in identifying dropout instances.

The F1 score for the negative class is 0.782, showing balanced performance.

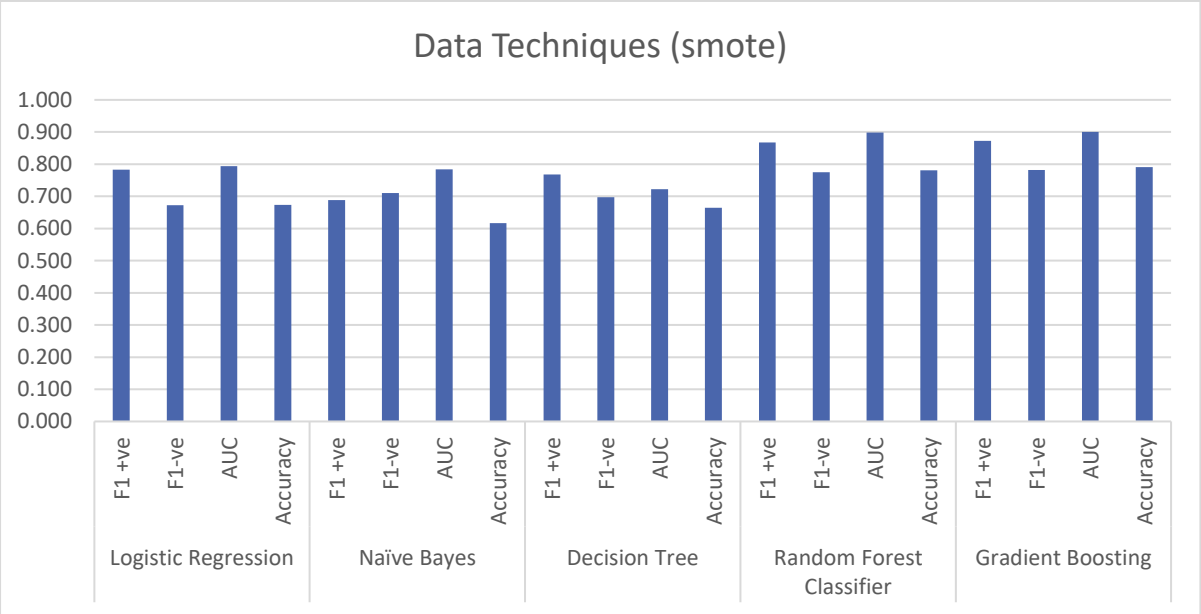
The AUC score of 0.900 suggests excellent discrimination ability, the highest among all models.

The overall accuracy increases to 0.791, indicating the effectiveness of Gradient Boosting in classifying dropout cases.

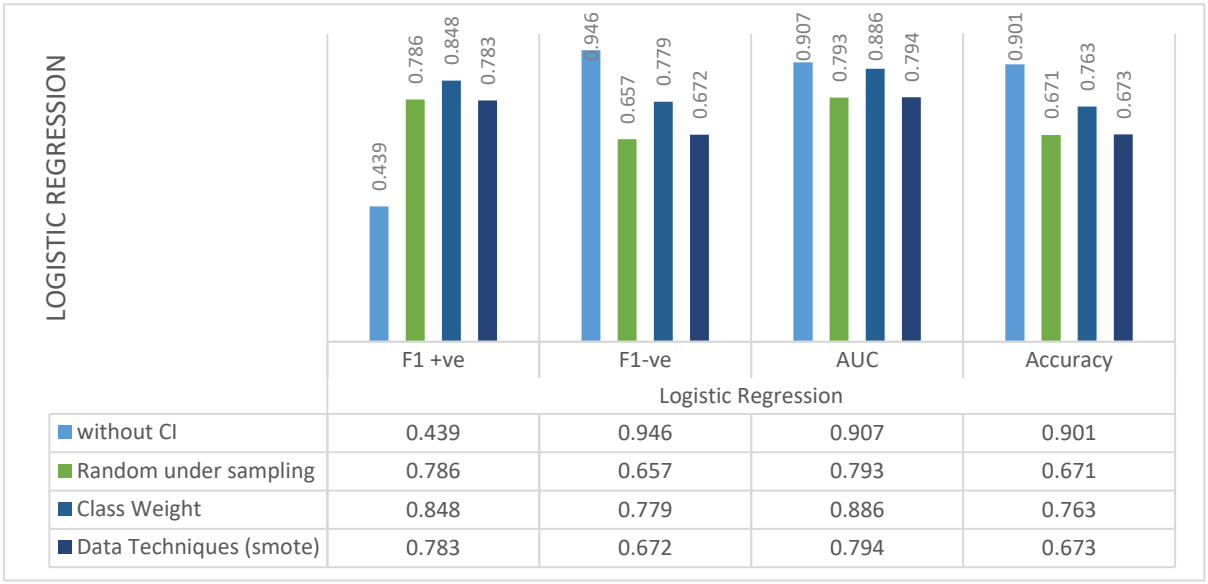
Conclusion:

Data techniques such as SMOTE improve the performance of all models in predicting dropout instances.

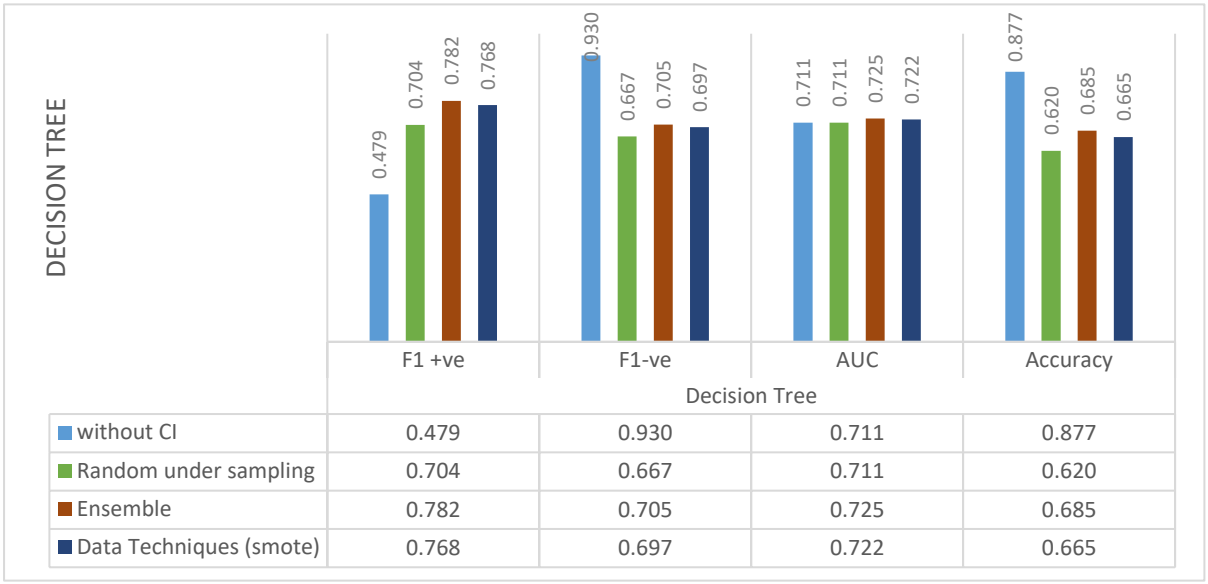
Gradient Boosting, in particular, shows significant improvement in identifying dropout cases with the highest F1 score, AUC, and accuracy among all models. This indicates its effectiveness in handling class imbalance and predicting dropout instances accurately.



Visual Representation of Combined Result:

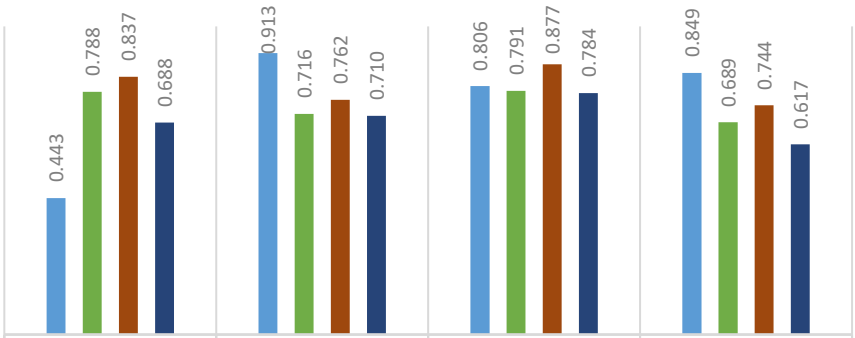


	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.439	0.946	0.907	0.901
Random under sampling	0.786	0.657	0.793	0.671
Class Weight	0.848	0.779	0.886	0.763
Data Techniques (smote)	0.783	0.672	0.794	0.673



	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.479	0.930	0.711	0.877
Random under sampling	0.704	0.667	0.711	0.620
Ensemble	0.782	0.705	0.725	0.685
Data Techniques (smote)	0.768	0.697	0.722	0.665

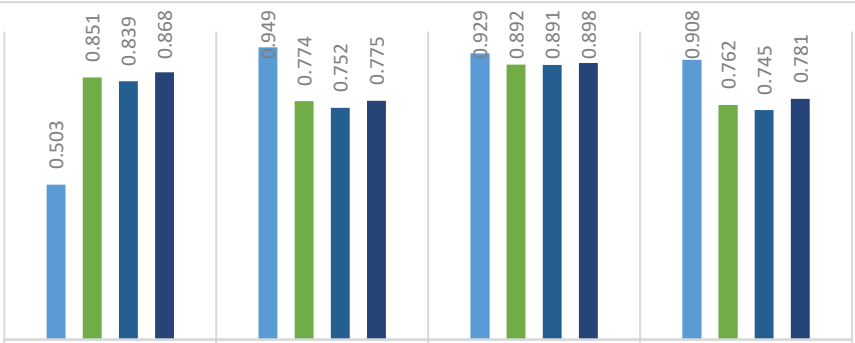
NAIVE BAYES



Naïve Bayes

	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.443	0.913	0.806	0.849
Random under sampling	0.788	0.716	0.791	0.689
Ensemble	0.837	0.762	0.877	0.744
Data Techniques (smote)	0.688	0.710	0.784	0.617

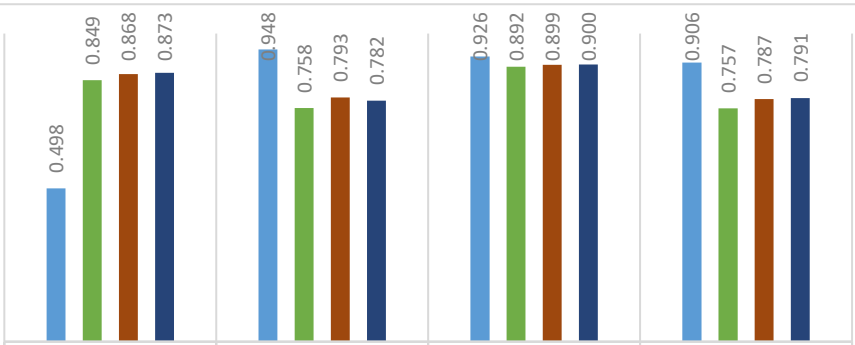
RANDOM FOREST



Random Forest Classifier

	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.503	0.949	0.929	0.908
Random under sampling	0.851	0.774	0.892	0.762
Class Weight	0.839	0.752	0.891	0.745
Data Techniques (smote)	0.868	0.775	0.898	0.781

GRADIENT BOOSTING



Gradient Boosting

	F1 +ve	F1-ve	AUC	Accuracy
without CI	0.498	0.948	0.926	0.906
Random under sampling	0.849	0.758	0.892	0.757
Ensemble	0.868	0.793	0.899	0.787
Data Techniques (smote)	0.873	0.782	0.900	0.791

Combined Analysis of Results:

Logistic Regression:

Without class imbalance, logistic regression performs poorly with an F1 score of 0.439 for the positive class and 0.946 for the negative class. The AUC is 0.907, indicating moderate discrimination ability.

Under random undersampling, the F1 score for the positive class improves significantly to 0.786, but the F1 score for the negative class decreases to 0.657. However, the AUC improves slightly to 0.793.

With class weight, there's a further improvement in the F1 score for the positive class (0.848) and a slight decrease in the F1 score for the negative class (0.779). The AUC improves to 0.886.

Using SMOTE, the F1 score for the positive class remains high at 0.783, and there's a slight improvement in the F1 score for the negative class to 0.672. The AUC also improves to 0.794.

Naïve Bayes:

Naïve Bayes shows consistent performance across different techniques. However, it performs poorly compared to other models.

The F1 score for the positive class ranges from 0.443 to 0.688, with the highest score achieved using SMOTE. The AUC ranges from 0.806 to 0.710.

Decision Tree:

The decision tree performs reasonably well in most scenarios.

Under random undersampling, it achieves an F1 score of 0.711 for the positive class and 0.716 for the negative class, with an AUC of 0.791.

With SMOTE, the F1 score for the positive class is 0.784, and the AUC is 0.784, indicating improved performance.

Random Forest Classifier:

Random forest shows good performance across all techniques.

Under random undersampling, it achieves an F1 score of 0.851 for the positive class and 0.774 for the negative class, with an AUC of 0.892.

With SMOTE, the F1 score for the positive class remains high at 0.868, and the AUC improves to 0.898.

Gradient Boosting:

Gradient boosting consistently outperforms other models across different techniques.

With SMOTE, it achieves the highest F1 score for the positive class (0.873) and the highest AUC (0.900), indicating excellent discrimination ability.

Conclusion:

Overall, gradient boosting with SMOTE stands out as the best-performing model for predicting dropout instances, demonstrating the importance of addressing class imbalance and utilizing advanced ensemble techniques for improved performance.

References:

https://archive.ics.uci.edu/dataset/222/bank+marketing

https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

https://machinelearningmastery.com/what-is-imbalanced-classification/

https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

https://medium.com/metaor-artificial-intelligence/solving-the-class-imbalance-problem-58cb926b5a0f

https://medium.com/@okanyenigun/handling-class-imbalance-in-machine-learning-cb1473e825ce

https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html

https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

https://chatgpt.com/c/bc28ec5b-1ada-43e8-b964-85550f571c08

==========

Thankyou