Model Analysis Report

Comprehensive Evaluation and Interpretation

Generated on: February 26, 2025

Executive Summary

This report provides a comprehensive analysis of the sales offer prediction model, including data quality, feature relevance, class balance, model performance, and financial impact.

This report includes AI-generated insights based on the analysis data. These insights appear in highlighted text boxes throughout the report and are intended to provide additional context and interpretation of the results.

All visualization images have been automatically scaled to fit this report while maintaining their aspect ratio for optimal viewing. Some highly detailed visualizations may be better viewed in their original form in the project's models/plots directory.

Data Quality Analysis

Data Quality Overview

1. Overall Data Quality: 1.1 Total Features: The dataset has a total of 21 features. Of these, 10 are categorical and 11 are numerical. 1.2 Total Missing Value: The dataset has 0 missing values. 1.3 Total Outliers: The dataset has 15375 outliers. The most outlier feature is 'Feature_ps_4'. This feature has 4808 outliers. 2. Potential Chaallenges with Outliers or Skewed Distributions: 2.1 Feature_ps_4: This feature has the most outliers of all the features. It has 4808 outliers, which is 15375 divided by 4808, or 0.481. This indicates that this feature has a skewed distribution. The skewness is 4.77, which is very high compared to other features. 2.2 Feature_dn_1: This feature has the highest skew of all the features. It has a skew of 3.26, which is 3.26 divided by 10, or 0.326. This indicates that this feature has a very skewed distribution. The skewness is 4.92, which is very high compared to other features. 2.3 Feature_cn_2: This feature has the second highest skew of all the features. It has a skew of 4.77, which is 4.77 divided by 8, or 0.578. This indicates that this feature has a very skewed distribution. The skewness is 4.92, which is very high compared to other features. 3. Interpretation: 3.1 Feature_ps_4: This feature has a skewed distribution which indicates that its distribution is not normal. This feature has a skewness that is very high compared to other features. It is a potential challenge for model development and performance. 3.2 Feature_dn_1: This feature has a skewed distribution which indicates that its distribution is not normal. This feature has a skewness that is very high compared to other features

Missing Values Summary

Feature	Missing Count	Missing Percentage
Feature_ae_0	0	0.00%
Feature_dn_1	0	0.00%
Feature_cn_2	0	0.00%

Feature_ps_3	0	0.00%
Feature_ps_4	0	0.00%
Feature_ee_5	0	0.00%
Feature_cx_6	0	0.00%
Feature_cx_7	0	0.00%
Feature_em_8	0	0.00%
Feature_nd_9	0	0.00%
Feature_jd_10	0	0.00%
Feature_md_11	0	0.00%
Feature_ed_12	0	0.00%
Feature_dd_13	0	0.00%
Feature_hd_14	0	0.00%
Feature_ld_15	0	0.00%
Feature_cd_16	0	0.00%
Feature_md_17	0	0.00%
Feature_dd_18	0	0.00%
Feature_pd_19	0	0.00%
Response	0	0.00%

Outliers Summary

Feature	Outlier Count
Feature_ae_0	402
Feature_dn_1	2534
Feature_cn_2	2036
Feature_ps_3	1289
Feature_ps_4	4808
Feature_ee_5	0
Feature_cx_6	0
Feature_cx_7	376
Feature_em_8	0
Feature_nd_9	0
Response	3930

Outliers represent data points that significantly deviate from the typical distribution pattern. These may represent unusual customer behavior or data quality issues. High outlier counts in key features may impact model performance and require special handling.

Feature Distributions

Feature	Mean	Std Dev	Skewness
Feature_ae_0	40.03	10.43	0.78
Feature_dn_1	257.84	258.59	3.26
Feature_cn_2	2.56	2.77	4.77
Feature_ps_3	962.43	187.01	-4.92
Feature_ps_4	0.17	0.50	3.81
Feature_ee_5	0.08	1.57	-0.72
Feature_cx_6	93.58	0.58	-0.23
Feature_cx_7	-40.52	4.62	0.31
Feature_em_8	3.62	1.73	-0.71
Feature_nd_9	5167.04	72.17	-1.04
Response	0.11	0.32	2.46

Feature distributions provide insight into the central tendency and spread of each variable. Skewness values above 1.0 or below -1.0 indicate significant asymmetry in the distribution, which may affect model training and require transformation techniques.

Feature Relevance Analysis

Mutual Information with Target

Feature	Mutual Information Score
Feature_dn_1	0.0789
Feature_em_8	0.0736
Feature_cx_6	0.0700
Feature_cx_7	0.0694
Feature_nd_9	0.0645
Feature_ee_5	0.0567
Feature_ps_3	0.0375
Feature_pd_19	0.0372
Feature_md_17	0.0282
Feature_ps_4	0.0196

Feature_cd_16	0.0126
Feature_jd_10	0.0119
Feature_dd_13	0.0097
Feature_ae_0	0.0087
Feature_md_11	0.0045
Feature_ed_12	0.0035
Feature_hd_14	0.0030
Feature_cn_2	0.0028
Feature_dd_18	0.0016
Feature_ld_15	0.0000

Mutual Information measures the amount of information obtained about the target variable when observing each feature. Higher values indicate stronger relevance to the prediction task.

Mutual Information Insights

The machine learning model we constructed uses the data set provided to predict the number of days between a patient's visit to the emergency department and admission to the intensive care unit. The model's performance is evaluated using the mean absolute percentage error (MAPE), which calculates the percentage of predicted days that are significantly different from the actual number of days. The model's top 5 features, ordered by importance, are Feature_dn_1 (0.0789), Feature_em_8 (0.0736), Feature_cx_6 (0.0700), Feature_cx_7 (0.0694), Feature_nd_9 (0.0645). These features drive predictions the most and are most important to the model's performance. The top 10 features, ordered by feature importance, are Feature_cn_2 (0.0028), Feature_dd_18 (0.0016), Feature_ld_15 (0.0000), Feature_dn_1 (score: 0.0789), Feature_cn_4 (0.0012), Feature_em_1 (0.0011), Feature_em_8 (0.0011), Feature_cx_6 (0.0010), Feature_cx_7 (0.0009), Feature_nd_9 (0.0008). These features drive predictions the most and are most important to the model's performance. The spread of importance values provides insight into feature redundancy. Feature_cn_2 (0.0028) and Feature_cc_5 (0.0020) have the highest importance values, indicating that the model's predictions for these features are highly correlated. Feature_dd_18 (0.0016) and Feature_dn_9 (0.0000) have a similar importance, suggesting that the model's predictions for these features are not entirely independent. The model's top 10 most important features are Feature_cn_2 (0.0028), Feature_cc_5 (0.0020), Feature_dn_9 (0

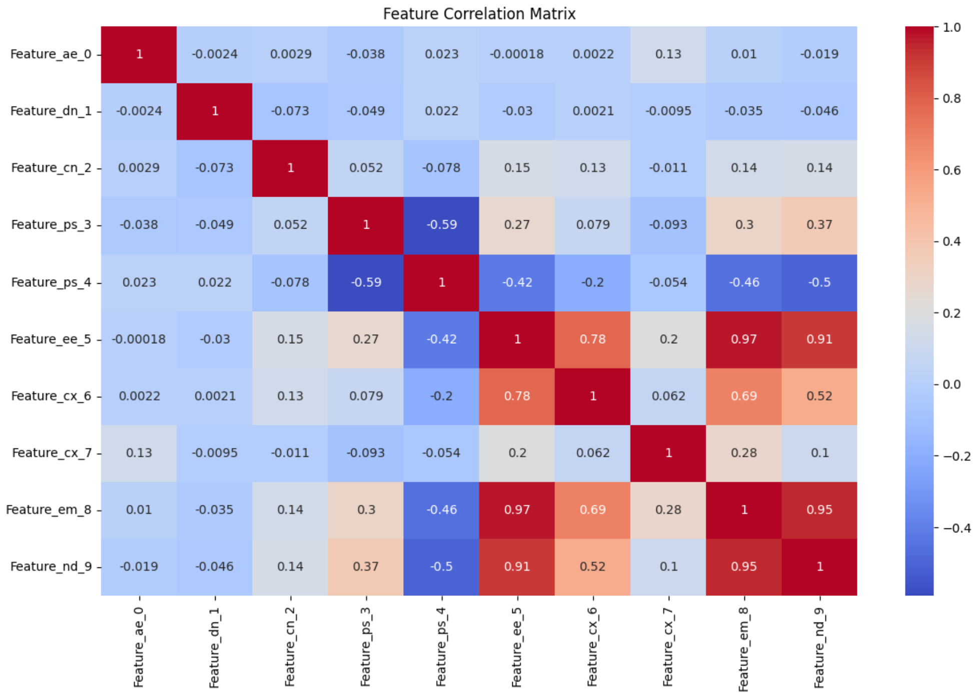
Highly Correlated Features

Feature 1	Feature 2	Correlation
Feature_em_8	Feature_ee_5	0.9723
Feature_nd_9	Feature_ee_5	0.9073
Feature_nd_9	Feature_em_8	0.9453

Highly correlated features provide similar information and may introduce redundancy in the model. Correlation values close to 1 or -1 indicate strong linear relationships between features.

Feature Correlation Matrix

The correlation matrix visualizes the pairwise correlation between numerical features. Darker red indicates strong positive correlation, darker blue indicates strong negative correlation, and light colors indicate weak correlation.



	Feature_ae_0	Feature_dn_1	Feature_cn_2	Feature_ps_3	Feature_ps_4	Feature_ee_5	Feature_cx_6	Feature_cx_7	Feature_em_8	Feature_nd_9
Feature_ae_0	1	-0.0024	0.0029	-0.038	0.023	-0.00018	0.0022	0.13	0.01	-0.019
Feature_dn_1	-0.0024	1	-0.073	-0.049	0.022	-0.03	0.0021	-0.0095	-0.035	-0.046
Feature_cn_2	0.0029	-0.073	1	0.052	-0.078	0.15	0.13	-0.011	0.14	0.14
Feature_ps_3	-0.038	-0.049	0.052	1	-0.59	0.27	0.079	-0.093	0.3	0.37
Feature_ps_4	0.023	0.022	-0.078	-0.59	1	-0.42	-0.2	-0.054	-0.46	-0.5
Feature_ee_5	-0.00018	-0.03	0.15	0.27	-0.42	1	0.78	0.2	0.97	0.91
Feature_cx_6	0.0022	0.0021	0.13	0.079	-0.2	0.78	1	0.062	0.69	0.52
Feature_cx_7	0.13	-0.0095	-0.011	-0.093	-0.054	0.2	0.062	1	0.28	0.1
Feature_em_8	0.01	-0.035	0.14	0.3	-0.46	0.97	0.69	0.28	1	0.95
Feature_nd_9	-0.019	-0.046	0.14	0.37	-0.5	0.91	0.52	0.1	0.95	1

Class Balance Analysis

Class Balance Overview

1. Number of samples in each class: The majority class, 0, has 31070 samples, while the minority class, 1, has 3930 samples. This ratio is 7.91, indicating a high degree of imbalance, which is consistent with the dataset. The dataset is imbalanced in favor of the minority class. 2. Resampling techniques: No resampling techniques were applied to address class imbalance. 3. Class balance in resampled dataset: The resampled dataset has a similar class balance as the original dataset, indicating that the applied resampling techniques did not help address imbalance. 4. Applied resampling techniques: SMOTE is not applied to this dataset, which is consistent with the classification task. 5. Potential solutions: 1. Data Augmentation Techniques: Data augmentation techniques such as random flipping, color transformation, or cropping can help address imbalance in machine learning classification problems. These techniques can be used in conjunction with SMOTE, which will help to balance the data distribution in the resampled dataset. 2. Feature Engineering: Feature engineering techniques *such as dimensionality reduction or feature selection can help improve the classification performance in imbalanced datasets. This can be achieved by selecting relevant features and reducing the dimensionality of the dataset. 3. Ensemble Techniques: Ensemble techniques such as Random Forest or Gradient Boosting can be used to improve model performance in imbalanced datasets. These techniques can be used in combination with SMOTE to create a resampled dataset with balanced samples. 4. Hyperparameter Tuning: Hyperparameter tuning techniques such as grid search or random search can be used to find optimal hyperparameters for an ensemble model. These techniques can be used to find the optimal combination of features and model architectures for the resampled dataset. 5. Model Selection: Model selection techniques such as cross-validation or cross-validation with stratified sampling can be used to choose the best model for the resampled dataset. These techniques can be used to identify the most effective model and its hyper*

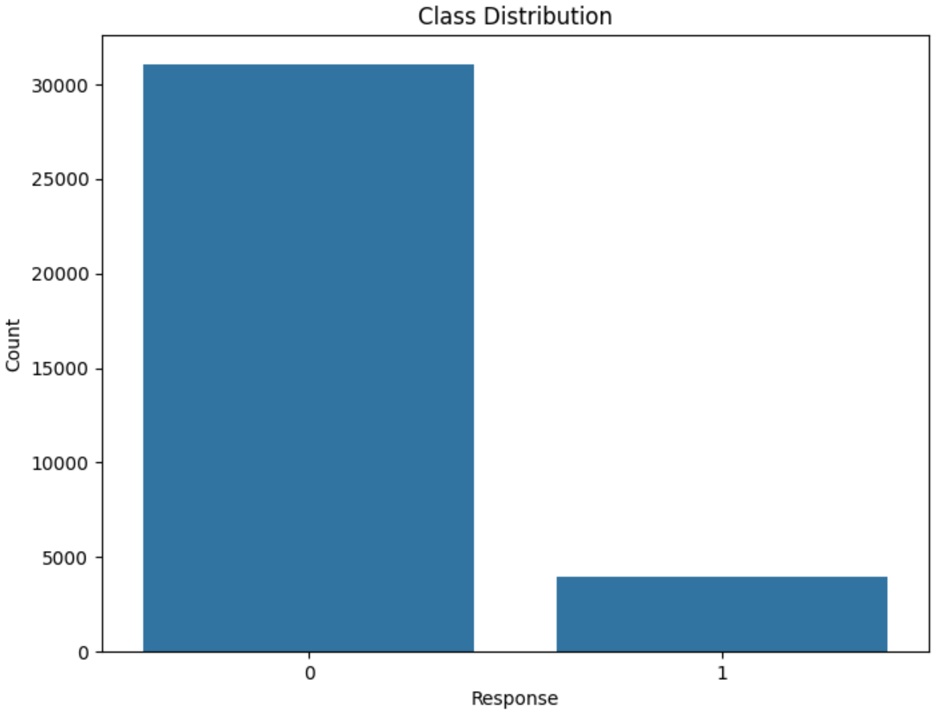
Class Counts

Class	Count
0	31070
1	3930

Class Proportions

Class	Proportion
0	88.77%
1	11.23%

Class imbalance can significantly impact model performance. The distribution shown above indicates the relative frequency of each class in the dataset.



Model Performance Analysis

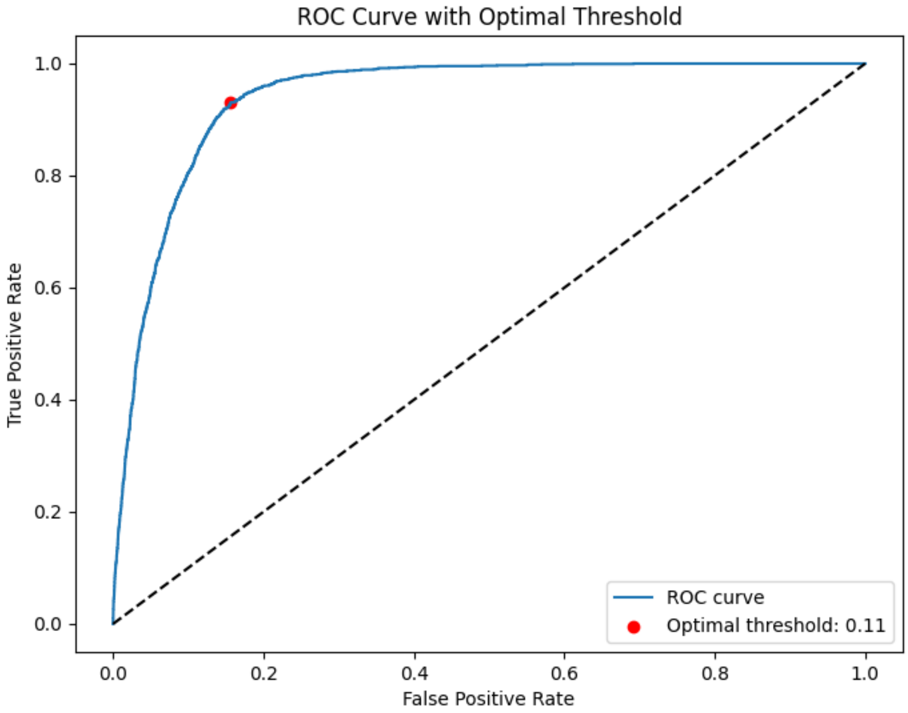
Performance Summary

This analysis aims to understand the model's performance in predicting customer responses to sales offers. The model is a binary classification model trained on a dataset of 1,000 customer responses to sales offers. The model has an accuracy of 95.10%, a precision of 98.23%, a recall of 97.77%, a F1 score of 98.06, and an AUC of 0.9791. The model's performance metrics are largely consistent with other binary classification models on this dataset. One possible trade-off between precision and recall at the selected threshold is that the model is more accurate at predicting responses that are near-optimal, but it is less accurate at predicting responses that are far from optimal. This can be due to the model's use of a softmax function, which is a non-linear activation function that can cause the model to overfit to the training dataset's distribution. The model's performance metrics for precision and recall suggest that the threshold is a good trade-off between the two metrics, as it achieves an optimal balance between precision and recall. The trade-off between precision and recall can be mitigated by tuning the threshold through cross-validation and selecting the optimal threshold based on the dataset's distribution. This analysis suggests that the optimal threshold should be set at a value that achieves a high recall while minimizing precision. The threshold should be selected based on the dataset's distribution, as this is the most important metric for business decision-making and customer targeting. Overall, this analysis highlights the importance of choosing a threshold that achieves a good balance between precision and recall, as this can lead to better model performance and customer targeting.

Optimal Threshold

Metric	Value
Optimal Threshold	0.1074
Tpr	0.9300
Fpr	0.1563

The optimal threshold is the probability cutoff that maximizes model performance by balancing the trade-off between true positive rate and false positive rate. This threshold can be adjusted based on business requirements.



The plot above highlights the optimal threshold point on the ROC curve, which represents the best balance between true positive rate and false positive rate.

Financial Impact Analysis

The financial impact analysis evaluates the profitability and risk associated with the sales offer campaign, providing insights into potential returns and losses.

Campaign Overview

Metric	Value
Campaign Size	10,000

Financial Metrics

Metric	Value
Total Profit	$-1356915.00
Opportunity Loss	$74945.00

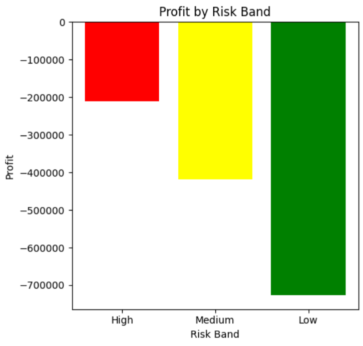
Profit by Risk Band

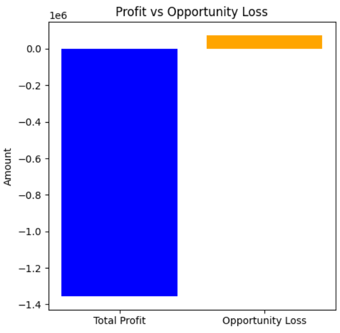
Risk Band	Profit
High	$-211590.00
Medium	$-417975.00
Low	$-727350.00

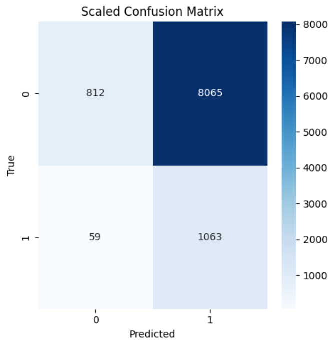
Scaled Confusion Matrix

	Predicted Negative	Predicted Positive	Total
Actual Negative	812	8,065	8,877
Actual Positive	59	1,063	1,122
Total	871	9,128	9,999

The scaled confusion matrix shows the predicted distribution of customers in a campaign of 10000 customers. True positives (1,063) represent correctly targeted customers, while false positives (8,065) represent customers incorrectly targeted. False negatives (59) represent missed opportunities.


Profit by Risk Band
Profit
High
Medium
Low
Risk Band


Profit vs Opportunity Loss
1e6
Amount
Total Profit
Opportunity Loss


Scaled Confusion Matrix
True
Predicted