Personalized Product Recommendation

Rishabh Gupta
rgupta416@gatech.edu

***Abstract*—** This report presents a predictive analytics model developed for NRG Energy, focusing on enhancing the adoption of its Ecoshare product. Using historical customer call data, the model predicts the likelihood of customers accepting Ecoshare, aiding agents in making precise product pitches. While centered on Ecoshare, this approach lays the foundation for personalized recommendations across NRG Energy's product range. The model employs advanced machine learning techniques and generates customer-specific propensity scores to guide tailored interactions. This streamlined strategy not only improves customer experience but also sets a scalable precedent for predictive analytics in customer relationship management and sustainable product promotion.

1 INTRODUCTION

NRG Energy, a leader in the energy sector, has been instrumental in offering innovative and environmentally friendly solutions, with Ecoshare as a standout product allowing customers to invest in carbon offsets (Reliant, n.d.). In the dynamic energy market, where customer preferences rapidly evolve, NRG Energy recognizes the need for a data-driven approach in understanding and influencing customer behaviors, especially concerning Ecoshare. The project at hand focuses on leveraging predictive analytics to enhance the company's customer engagement strategies, particularly in the promotion and adoption of Ecoshare.

The core objective of this project is to develop a predictive model that predicts the likelihood of customers adopting Ecoshare. This model aims to improve the

interaction strategies at NRG Energy's customer retention center by analyzing historical call data to refine the techniques agents use in pitching Ecoshare. While the immediate focus is on Ecoshare, the broader aim is to lay the groundwork for a system of personalized product recommendations across NRG Energy's product range, enhancing customer relationships.

Advanced machine learning algorithms are employed to analyze customer interaction data, with a significant focus on creating customer-specific propensity scores. These scores are designed to predict customer interest in Ecoshare and assist agents in making targeted product recommendations. The successful implementation of this model is expected to not only improve the accuracy of product pitches but also to increase customer satisfaction through more personalized service offerings, potentially leading to higher adoption rates of Ecoshare. Additionally, the insights and methods derived from this project are envisioned to be applicable to other products, extending the model's impact across NRG Energy's portfolio.

This report will detail the predictive model's development process, its application in real-time customer interactions, and its potential for broader application across different products. The project signifies a significant stride in integrating advanced data analytics into NRG Energy’s customer engagement strategies, setting new standards in the energy sector for combining customer-centric services with a commitment to environmental sustainability.

2 DATA AND PREPROCESSING

2.1 Data

The data utilized for this predictive analytics project was provided by the NRG Team and is centered on customer interactions relating to the Ecoshare product. This dataset captures a comprehensive snapshot of customer behaviors and preferences during their engagements with the customer retention center. The period from January 2017 to December 2020 served as the training data timeframe,

with the subsequent period up to September 2023 used for testing, reflecting the interactions of approximately 52,000 unique customers.

The dataset is rich with 33 distinct features that shed light on various aspects of customer data. These features are segmented into several categories:

- Customer Interaction and Service History: This category includes critical data points such as the date when the Ecoshare offer was made, the customer's history of service transfers, any received disconnect notices, and the frequency of interactions with the call center.
- Customer Account and Usage Information: Here, the dataset covers the length of the customer's electricity contract, recent electricity usage, the type of contract they hold, and metrics pertaining to online engagement with the service provider.
- Risk and Payment Profiles: This section delves into the risk level associated with the customer's payment of bills, any history of late fees, and the chosen payment methods.
- Demographics and Home Details: These features encompass demographic information, the estimated value of the customer's home, geographical data including city, county, and zipcode, and details such as whether the customer's residence has a swimming pool.

At the heart of this analysis is the binary target variable 'accept', which indicates whether a customer agreed to the Ecoshare offer during the interaction. This variable shows an acceptance rate of about 17% across customer interactions, it presents a significant challenge due to its imbalanced nature. (Refer to Figure 1 for a visualization of the 'accept' variable distribution.)

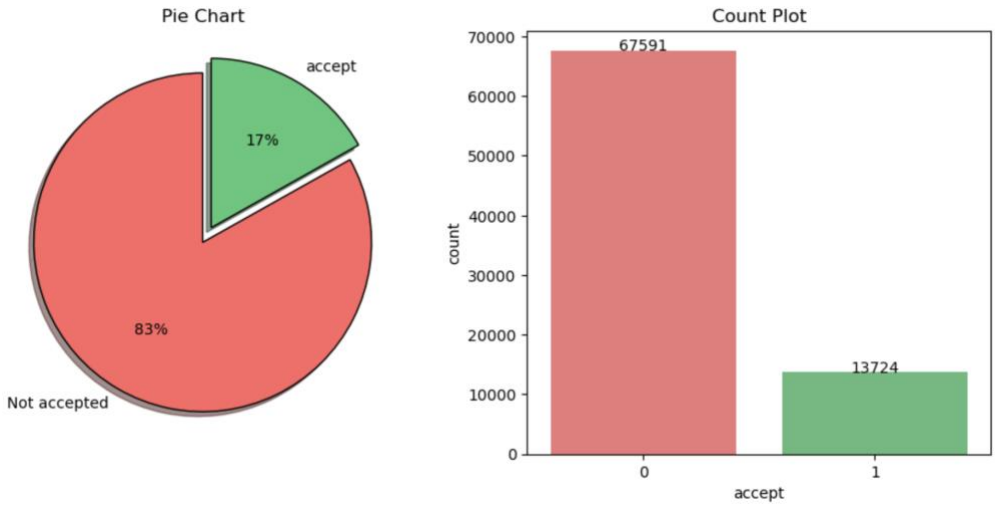

Figure 1: Distribution of NRG Energy customer has accepted EcoShare.

2.2 Preprocessing

In the data preprocessing phase of the project, a structured approach was adopted to address missing values and transform the data for optimal use in predictive modeling. The process began with the imputation of missing values in the 'tos_flg' feature, which was set to 'N', indicating no service transfer where data was absent. For the 'curr_usage' and 'deposit_onhand_amt' features, missing values were replaced with zeros, operating under the assumption that a lack of data signified non-usage or the absence of a deposit. To maintain consistency in the categorical variables 'risk_level' and 'sap_productname', missing values were filled with the most common value, or the mode, of the respective features.

Further into the preprocessing stage, certain features underwent transformation to better fit the requirements of the predictive models. The 'zipcode' and 'segment' features, for example, were converted to object types due to their categorical nature. From the 'order_day' feature, additional date-related features such as 'orderyear', 'Ordermonth', and 'Isweekend' were extracted. These new variables aimed to capture the temporal dynamics of customer interactions, which could be pivotal in understanding customer behavior patterns that are influenced by time,

such as seasonality and weekends. Correlation between numerical features is shown in in Figure 2

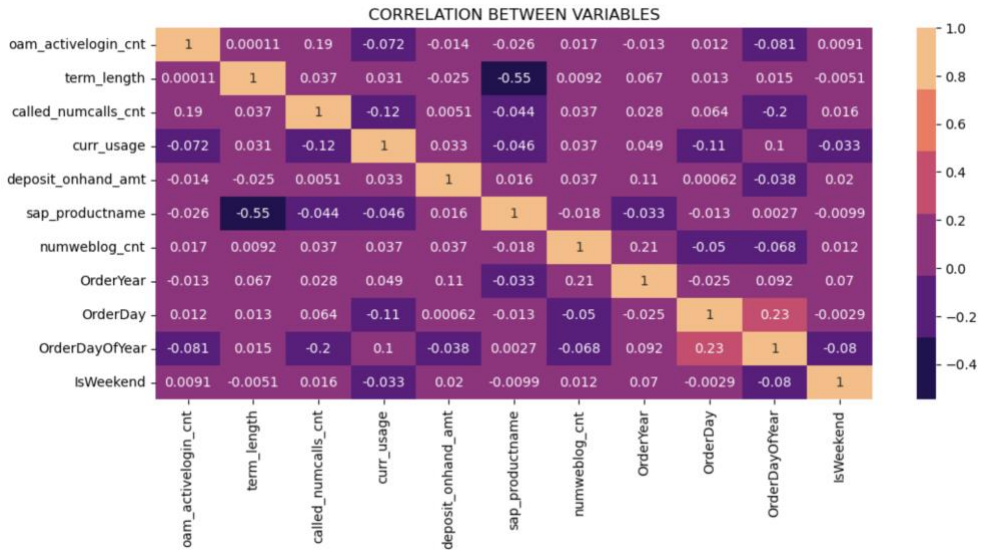

Figure 2: Correlation Plot

However, not all features were retained. Due to a high rate of missing values, features like 'pool' and 'home_value' were excluded from the dataset, as visualized in Figure 3. Their removal was crucial to prevent potential bias and inaccuracies within the predictive model. For the 'sap_productname', frequency encoding was applied, replacing each category with its frequency count within the dataset. This method was chosen to reduce the complexity that would result from one-hot encoding and to preserve the essential information about the frequency of each category.

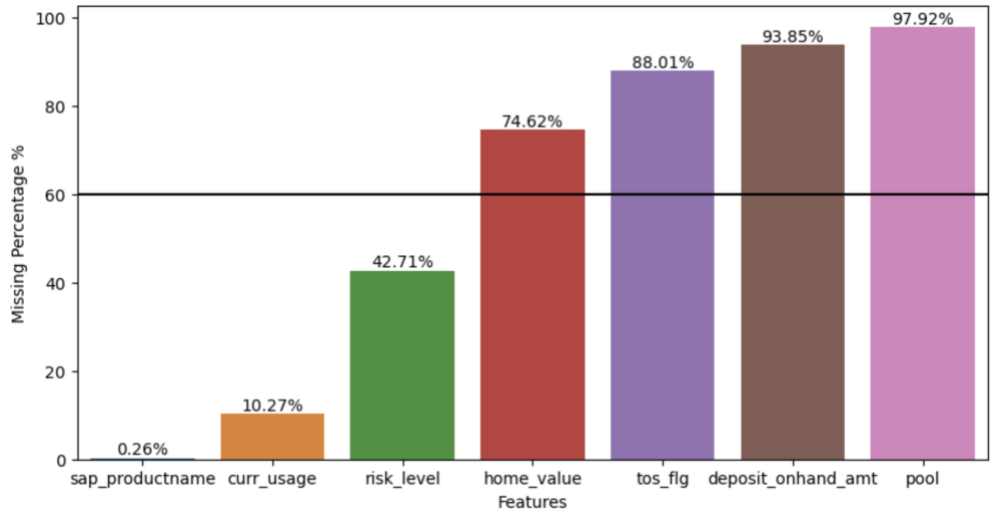

Figure 3: Missing values

Geographic features such as 'city', 'zipcode', 'county', and 'dma' underwent a consolidation process where only the top three categories were maintained, and the remaining were grouped into an 'other' category. This approach was designed to streamline the model by focusing on the most impactful categories and reducing the risk of overfitting that can arise from an excessive number of categories.

For the categorical features, ordinal encoding was utilized to convert them into a format that could be interpreted by the machine learning algorithms, preserving any inherent order within the categories. This step was essential for the tree-based models, which are adept at handling categorical variables. For non-tree-based models, which typically benefit from more delineated categorical variables, one-hot encoding was also applied in addition to ordinal encoding. Numerical features were normalized using the MinMax scaling technique, essential for standardizing the input features' range and particularly beneficial for models sensitive to the scale of data.

A significant preprocessing step involved addressing the class imbalance in the target variable, 'accept'. Class imbalance can lead to a model's predictive performance bias towards the majority class, thereby hindering its ability to

accurately predict the minority class. To counteract this, two primary techniques were utilized:

Resampling Techniques:

- UnderSampling: Reducing the number of instances from the majority class to balance the dataset.
- OverSampling: Increasing the number of instances in the minority class.
- SMOTE (Synthetic Minority Over-sampling Technique): Generating synthetic examples of the minority class to achieve balance (SATPATHY, 2020).

Additionally, variants of SMOTE were explored:

- SMOTEENN (SMOTE + Edited Nearest Neighbors): This technique combines SMOTE with the ENN algorithm, which cleans the dataset by removing any instance of the majority class that has a majority of neighbors from the minority class.
- SMOTETomek (SMOTE + Tomek Links): It pairs SMOTE with Tomek Links, where Tomek links are pairs of very close instances but of opposite classes. Removing the instances of the majority class from these pairs can help in cleaning overlapping points between classes.

Other oversampling techniques such as Nearmiss, a version of undersampling that selects instances of the majority class closest to the minority class boundary, were also considered (Brownlee, Tour of Data Sampling Methods for Imbalanced Classification, 2020). This can help in creating a more defined decision boundary for the model.

Algorithmic Ensemble Techniques:

- Bagging: Combining the predictions of multiple models to reduce variance and improve results.
- Boosting: Sequentially building models that correct the errors of the models before them.
- Class Weights: Assigning a higher weight to the minority class during model training, often by setting class_weights='balanced' (Brownlee, Cost-Sensitive Logistic Regression for Imbalanced Classification, 2020).

These techniques were chosen to enhance the model's ability to predict customer adoption more accurately, particularly for customers likely to accept the Ecoshare offer but who are underrepresented in the dataset. By ensuring a more balanced class distribution, the predictive models were equipped to offer a more nuanced and equitable assessment of the likelihood of customer adoption, which is crucial for developing effective customer engagement strategies.

The preprocessing strategies, including feature transformation and class imbalance mitigation, were integral to preparing the dataset for robust and reliable predictive modeling. These efforts set the stage for deploying advanced algorithms capable of delivering actionable insights into customer behavior.

3 METHODOLOGY

In this project, a comprehensive approach was adopted to develop a predictive model for estimating the probability of customers adopting NRG Energy's Ecoshare product. The methodology encompassed the careful selection of machine learning algorithms, thorough data preparation, rigorous training and validation processes, and a detailed evaluation of the model's performance.

Following the principles of the "No Free Lunch" theorem, which suggests that no single algorithm excels in all predictive tasks, a diverse array of both tree-based (such as Random Forest and Gradient Boosting) and non-tree-based models (like Logistic Regression, Naïve Bayes and Support Vector Machines) were selected (Mavuduru, 2020). This varied choice of algorithms was aimed at ensuring a well-rounded analysis of the dataset. To refine these models, Bayesian optimization was employed for hyperparameter tuning, an approach renowned for its efficiency in balancing the exploration and exploitation of the search space, thereby optimizing model performance (BANERJEE, 2020).

The dataset was split into training and test sets, allocating 20% of the training data for testing using stratified sampling. This method ensures that the test set represents the overall dataset, providing a realistic view of how the model would perform in a production environment. To further assess the model's generalizability and prevent overfitting, a 5-fold cross-validation strategy was implemented.

Different sets of features were prepared for the tree-based and non-tree-based models. While ordinal encoding was applied to the features for the tree-based models, one-hot encoding was used for the non-tree-based models. In addition, SHAP (SHapley Additive exPlanations) values were utilized to analyze the importance of different features in the tree-based models, offering a deeper understanding of how each feature influences the model's predictions (Trevisan, 2022).

The evaluation of the model was primarily focused on AUC-ROC and Log Loss metrics, considering their relevance for models predicting probabilities and their independence from specific decision thresholds. This focus allowed for an objective comparison of different models based on their ability to predict customer adoption of Ecoshare accurately.

A significant emphasis was also placed on ensuring the interpretability of the model outputs. The ability to clearly understand and communicate the influence of various features on the model's predictions is essential, especially for translating these insights into actionable strategies for business stakeholders.

This rigorous and systematic approach outlines our first strategy in the development of the predictive model. Advanced techniques like Bayesian optimization and SHAP values were integrated, reinforcing our commitment to creating a robust, efficient, and transparent model.

Following this initial approach, we also explored alternative strategies to address the imbalance in the dataset, which is a common challenge in predictive modeling. We investigated resampling techniques, including UnderSampling, OverSampling, and advanced methods like SMOTE, SMOTEENN, and SMOTETomek, to achieve a more balanced class distribution. Additionally, Algorithmic Ensemble Techniques such as Bagging, Boosting, and utilizing class_weights='balanced' were applied. These methods aimed to enhance the model's performance, particularly for the minority class, which is often underrepresented in the data.

Moreover, the precision-recall curve was introduced as an additional metric for evaluation. This metric is particularly informative when dealing with imbalanced datasets, as it focuses on the performance of the minority class by illustrating the trade-off between precision and recall for different threshold settings. The precision-recall curve complements the AUC-ROC by providing a more detailed view of the model's capability in distinguishing the class of interest.

The integration of these resampling techniques, ensemble methods, and additional evaluation metrics allowed us to adopt a more nuanced approach in our second strategy, aiming to further refine the model's accuracy and ensure its efficacy in a practical, operational context.

4 Model Evaluation and Results

Addressing class imbalance is often a significant step in the modeling process, particularly when predicting outcomes for a minority class. Various techniques were applied to tackle this issue, including resampling methods like UnderSampling, OverSampling, SMOTE, and its more nuanced forms,

SMOTEENN and SMOTETomek. Algorithmic ensemble techniques such as Bagging, Boosting, and the use of class_weights='balanced' were also utilized. However, these techniques did not lead to an improvement in model performance, as indicated by Table 1. This finding suggests that for this specific dataset and prediction task, class imbalance strategies did not enhance the model's predictive accuracy.

Sampling Type	Model	AUC_ROC	F1	Precision	Recall
RandomUnderSampler	RandomForest	0.72	0.57	0.70	0.47
base	RandomForest	0.72	0.56	0.69	0.47
NearMiss	RandomForest	0.71	0.56	0.70	0.47
RandomOverSampler	RandomForest	0.71	0.56	0.69	0.47
SMOTETomek	RandomForest	0.71	0.56	0.69	0.47
SMOTEENN	RandomForest	0.71	0.56	0.69	0.47
SMOTE	RandomForest	0.71	0.56	0.69	0.47
base	LogisticRegression	0.69	0.51	0.60	0.45
SMOTE	LogisticRegression	0.69	0.51	0.60	0.45
RandomOverSampler	LogisticRegression	0.69	0.51	0.60	0.45

Table 1: Model Results after applying different Sampling Type

Without applying any of the class imbalance techniques, the results of the 5 fold cross validations mean were as follows: XGBoost outperformed other models with high scores across all metrics, showing particular strength in AUC-ROC (0.94) and LogLoss (0.21). The model's balance between precision and recall is noted in its F1 score (0.69), suggesting it can accurately identify potential adopters. XGBoost's efficiency was underscored by its low fitting and scoring times, emphasizing its suitability for handling diverse data types and classification tasks.

The Random Forest model showed similar proficiency, slightly trailing XGBoost in AUC-ROC (0.94) but with a slightly higher precision at the expense of recall. Logistic Regression and Support Vector Machine, despite being faster to fit, were less effective according to predictive metrics such as AUC-ROC and LogLoss. CatBoost, while scoring quickly, did not offer LogLoss performance on par with the leading models, and Naive Bayes, although quick computationally, exhibited considerable precision deficiencies and an inflated LogLoss.

When the final XGBoost model was tested on an unseen dataset, it achieved an AUC score of 0.81, as showcased in Figure 4. This confirmed the model's generalizability and its ability to discriminate effectively between adopters and non-adopters.

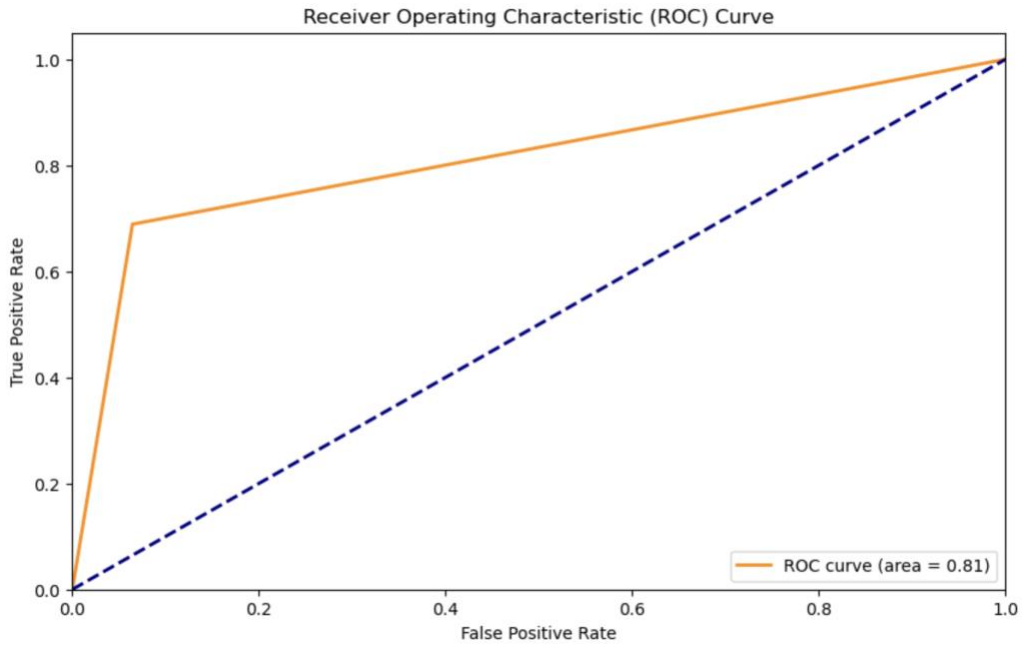

Figure 4: AUC ROC Curve for XgBoost Model

In summary, the ensemble methods, particularly XGBoost, were superior for this predictive task, demonstrating a potent combination of accuracy and computational practicality. These methods proved adept at navigating the complex patterns in the dataset, thereby providing robust predictions for customer adoption behavior.

The results of the model evaluation are summarized (As in Table 2), which serves as a reference for comparing the performance of the evaluated models. The investigation into class imbalance techniques did not yield a marked improvement in the model's performance. The initial results without these techniques already provided robust predictive accuracy. The experience gathered from this analysis reaffirms that while class imbalance strategies are valuable tools in predictive modeling, their applicability and impact can vary greatly depending on the dataset and the specificities of the prediction task at hand. This underscores the importance of a tailored approach to predictive modeling, where standard practices may need to be adapted or even set aside based on the unique dynamics of each project.

Model	AUC_ROC	LogLoss	F1_score	Fitting time (Seconds)	Scoring time (Seconds)
XGBoost	0.95	-0.21	0.69	1.98	0.03
Random Forest	0.95	-0.22	0.67	2.93	0.30
Logistic Regression	0.89	-0.29	0.52	1.14	0.03
Support Vector Machine	0.88	-0.32	0.39	7.37	1.75
CatBoost	0.92	-0.37	0.54	0.07	0.02
Naive Bayes	0.81	-4.49	0.43	0.04	0.02

Table 2: Model Results

The selection of the appropriate model was guided not only by the highest statistical measures but also by the model's fit for operational deployment, considering both the fitting and scoring times. XGBoost stood out as the model that best met these criteria, followed closely by Random Forest. These findings underscore the value of ensemble techniques in predictive analytics, particularly for applications in customer behavior prediction.

5 Feature Importance & Interpretation

XGBoost Model Insights

Our predictive analysis using XGBoost has highlighted 'curr_usage', 'called_flg', 'called_numcalls_cnt', 'OrderDayOfYear', 'product_type_cd', 'OrderYear', and 'OrderMonth' as the top seven features with the most substantial impact on the model's predictions, as demonstrated in Figure 5. These features were pivotal in determining the likelihood of a customer adopting EcoShare.

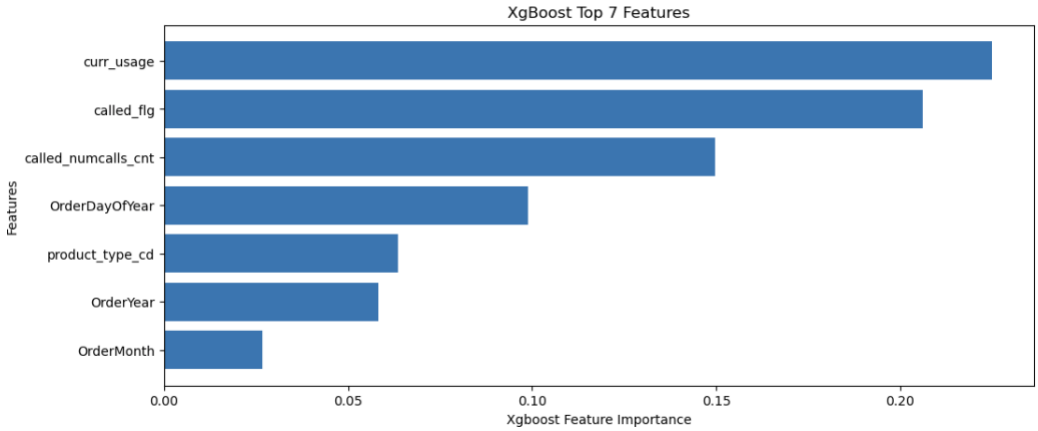

Figure 5: Feature Importance

SHAP Value Analysis of Feature Impact

The SHAP value analysis provides an in-depth look at the effect of each feature on the model's output, offering a nuanced perspective beyond what is typically captured by feature importance rankings alone (Figure 6).

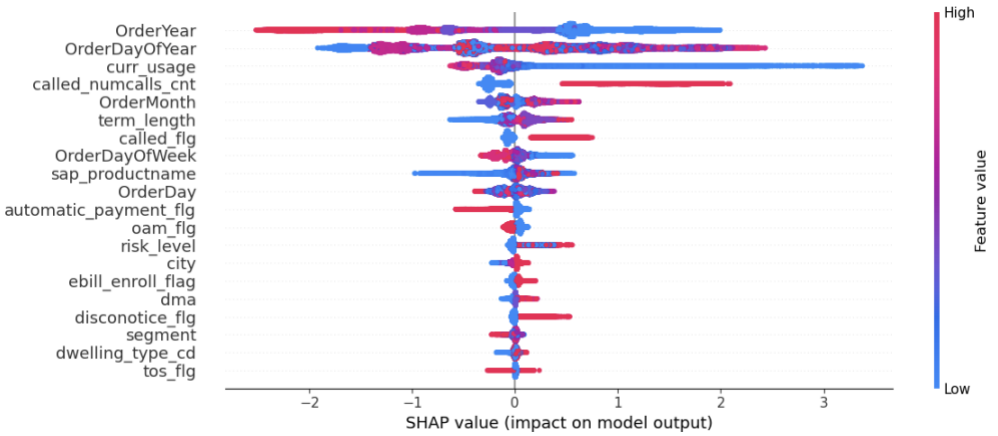

Figure 6: SHAP values Feature Importance

- Order Year ('OrderYear'): The SHAP values indicate that higher 'OrderYear' values have a significant positive impact on the model's predictions. This suggests that customers are more likely to engage with EcoShare in more recent years, which could be attributed to evolving consumer awareness or improved service offerings.
- Current Usage ('curr_usage'): Interestingly, lower 'curr_usage' values are associated with a higher impact on the model's output. This may imply that customers with lower electricity consumption are surprisingly more inclined to adopt energy efficiency programs like EcoShare, possibly as a cost-saving measure.
- Number of Calls ('called_numcalls_cnt'): High 'called_numcalls_cnt' values have a pronounced positive effect on the model's output. This aligns with the expectation that increased customer service contact may reflect a

higher level of customer engagement and a greater probability of adopting new services.

The SHAP value analysis extends our understanding of feature influence and provides a clearer direction for strategic decision-making. For instance, the impact of 'OrderYear' suggests that marketing strategies should evolve to capitalize on the increasing trend of adoption in recent years. The counterintuitive finding regarding 'curr_usage' could lead to targeted marketing efforts aimed at customers with lower usage levels, who may be looking for cost efficiency and sustainability. Lastly, the significance of 'called_numcalls_cnt' reinforces the value of customer engagement initiatives in driving service adoption.

The dual analysis using XGBoost feature importance and SHAP values offers a comprehensive understanding of the driving factors behind customer decisions to adopt EcoShare. By considering the magnitude and direction of SHAP values, we gain actionable insights that can enhance customer targeting, improve service offerings, and ultimately drive the adoption of EcoShare. This analytical approach ensures that our predictive modeling translates into effective business strategies.

6 RECOMMENDATIONS & FUTURE WORK

Personalized Product Recommendation Strategies

The predictive model developed for NRG Energy has demonstrated significant potential in influencing customer decisions regarding the adoption of the EcoShare product. Moving forward, the following strategies are recommended to leverage the model's capabilities and extend its application:

1. Expand Personalization Techniques: Continue to refine the predictive model to enhance its precision in scoring customer propensity. This could involve integrating more customer-specific data points or developing more

complex algorithms that can better handle the nuances of individual customer behaviors.

2. Broader Product Application: Consider developing a multiclass classification model to predict customer propensities across all products within a single dataset. This unified approach ensures comparability of probability scores, offers a more comprehensive view of customer preferences, and enhances operational efficiency. It also opens avenues for effective cross-selling by identifying which customers are likely to be interested in multiple products.
3. Customer Feedback Integration: Incorporate customer feedback mechanisms to continuously improve the recommendation system. Understanding customer perceptions of EcoShare and other products can provide valuable insights for model refinement.

Future Research Directions

To build on the current model and address identified challenges, the following areas of research should be pursued:

1. Algorithmic Diversity: Investigate additional machine learning algorithms and ensemble methods that might offer improved performance or interpretability, particularly in the context of class imbalance.
2. Feature Exploration: Conduct further research into the features that influence customer adoption, particularly examining the impact of economic and environmental trends on customer behavior.
3. Real-time Analytics: Assess the feasibility of implementing real-time analytics to provide agents with instantaneous recommendations during customer calls, further personalizing the customer experience.

7 CONCLUSIONS

The project concluded with several key findings: The XGBoost model emerged as the superior predictive tool, showcasing high AUC-ROC (0.81) , which highlighted its ability to accurately identify potential adopters of the EcoShare product. Efforts to mitigate class imbalance through various resampling methods did not significantly enhance the model's performance, indicating the robustness of the initial model. Feature importance analysis pinpointed 'curr_usage', 'called_flg', 'called_numcalls_cnt', 'OrderDayOfYear', 'product_type_cd', 'OrderYear', and 'OrderMonth' as critical predictors, with SHAP value analysis affirming their substantial impact on model predictions. Notably, more recent years, lower electricity usage, and frequent customer calls were identified as key drivers of likelihood to adopt EcoShare. The model also demonstrated strong generalizability when applied to unseen data, confirming its reliability for practical use. These insights provide actionable guidance for NRG Energy's marketing and customer engagement strategies, suggesting a focus on recent interactions and customers with lower usage levels to predict product adoption effectively. Overall, the project highlights the valuable role of predictive analytics in refining customer relationship management and promoting sustainable product options.

REFERENCES

(n.d.). Retrieved from Reliant: https://www.reliant.com/en/residential/home-solutions/renewable-products/reliant-ecoshare-program

BANERJEE, P. (2020). *Bayesian Optimization using Hyperopt* . Retrieved from Kaggle: https://www.kaggle.com/code/prashant111/bayesian-optimization-using-hyperopt

Brownlee, J. (2020, October). *Cost-Sensitive Logistic Regression for Imbalanced Classification*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/cost-sensitive-logistic-regression/

Brownlee, J. (2020, Januray). *Tour of Data Sampling Methods for Imbalanced Classification*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/

Mavuduru, A. (2020, November). *What "no free lunch" really means in machine learning*. Retrieved from towardsdatascience: https://towardsdatascience.com/what-no-free-lunch-really-means-in-machine-learning-85493215625d

SATPATHY, S. (2020, October). *SMOTE for Imbalanced Classification with Python* . Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Trevisan, V. (2022, January 18). *Using SHAP Values to Explain How Your Machine Learning Model Works*. Retrieved from Towards Data Science: https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137