Applied Analytics Practicum
Final Report

Deborah Chang
Sponsor: NRG Energy

Table of Contents

Abstract
This paper discusses the practicum completed with NRG Energy Inc., including the objective, dataset preparation and cleaning, modeling, and results. The goal was to develop a model that would predict whether one of their products, Reliant EcoShare, would be accepted during a given customer support call. This is part of a larger effort on providing insights at the company to leverage analytics methods to recommend products to customers. The dataset was historical call center data and was analyzed, cleaned, and leveraged to be modeled. The data was cleaned and then used for extra trees, random forest, and logistic regression models. Performances were calculated and then the best model was run on the test data to provide the probabilities of EcoShare being accepted on a call.

During the middle of the semester, the data needed to be recut twice due to potential leakage issues, so the insights and models that were initially derived have evolved but not significantly.

Overall, the experience provided me with significant learnings including how to navigate imperfect datasets, work with assumptions, learn the context behind the data, and how to derive valuable insights and recommendations for the business.

About the Sponsor
I chose to do a Georgia Tech sponsored project and partnered with NRG Energy Inc. NRG Energy is a Fortune 200 company that serves electricity, natural gas, and smart home products to homes and businesses across North America. Examples of products are electricity and solar plans, EV driving, smart home systems for automated security, and protection plans for costly repairs. The company also focuses on sustainability and social impact to power positive change on people, communities, and the environment. More information can be found at nrg.com. Here is a press release on the latest acquisition. As quoted, "The combined company will be the leading essential home solutions provider, with an extensive network of approximately 7.4 million customers across North America, that represents a substantial cross-sell opportunity through market-leading brands and complementary sales channels."

Reliant Ecoshare Program
The key product I looked into during the practicum was NRG's Reliant EcoShare Program. Reliant was acquired back in 2009. The EcoShare program goes towards carbon offsets, which is an effort to compensate for carbon emission such as waste management or reforestation and towards clean energy efforts. A small charge gets added to the customer's monthly bill - a portion goes to an environmental nonprofit called EarthShare and the rest goes toward carbon offsets. There are two plans - Reliant EcoShare Gold and Reliant EcoShare Silver, where Gold has a higher pricing but more offsets and donations to EarthShare of Texas. More information on the program can be found here.

Objective
Given retention call center data, the objective was to develop a data model that will predict whether a Reliant Energy customer will accept participating in the EcoShare program when a Customer Care agent pitches the offer during a call. The model will be applied as part of a larger effort in NRG's Cross Serve Personalization Engine that will recommend useful and relevant products for homes. The main deliverable is to provide the probabilities of acceptance for the test

set that is provided.

Business Impact of Retention Call Analysis

Many companies focus on how call center data can reveal many insights on customer patterns and what products they may like, their behaviors and interactions with the agents, their buying committee, common troubleshooting questions, and more. Analytics of call center data can also provide paths to reduce call center costs and give insights on products that customers may be more likely to purchase. As stated in Invoca's blog, "How to Analyze Call Center Data to Improve Efficiency," retention call analysis can provide impressive insights on ROI. That is, if they are backed by call center improvement strategies that help drive efficiency and improve productivity, while also enhancing the customer experience (Andersen).

Retention call center data can also provide invaluable insights on customer service, attrition, and the operational processes that go into it (Call Center Administration, 2023). Feedback can also be collected through surveys and qualitative comments from agents. It can help to see whether more guidance or additional agent training is needed.

The Data

The training dataset has around 81K records of historical retention call center data from January 2017 to December 2020 between a Customer Care agent and a Reliant Energy customer where EcoShare was most likely pitched to the customer. Features include product name, order date, term length, home features like home value and whether there is a pool, website activity, payment delay and method, electricity consumed, and location. See the Appendix section for the full data dictionary. These are some caveats as stated by the sponsor:

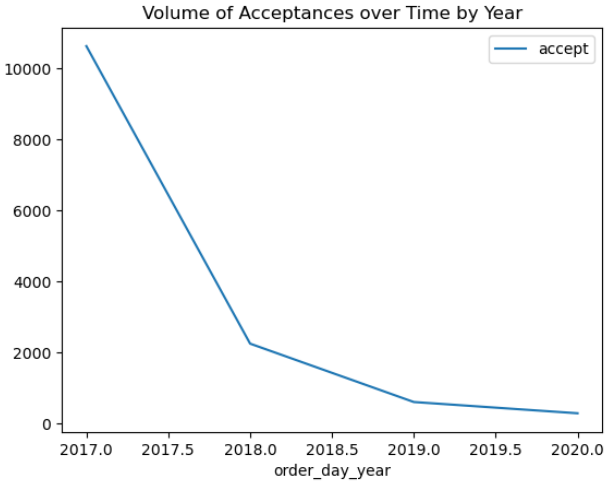
- For each call, the customer is eligible for the EcoShare program
- There may be multiple calls for a customer e.g. they have multiple products or properties
- It is most likely that EcoShare was offered - it is one of the top 2 products to be pitched, so the agent could have proposed EcoShare, the other product, or both
- There are cases where EcoShare was not offered due to agent inexperience or billing issues that take up the whole duration of the call

Towards the latter half of the course, due to data leakage issues, the training data had to be recut a few times to ensure that the timeframe was appropriate for modeling and performance testing.

Exploratory Data Analysis

I looked into the features below to see whether there were any significant trends or patterns for customers who accepted EcoShare on a given call versus customers who did not. Some of the fields were transformed before being analyzed (e.g. weblog_flg, term_length).

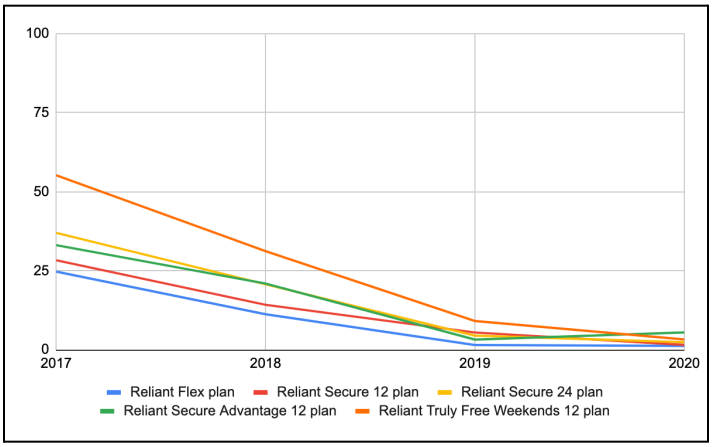
Acceptance Trends



Almost 17% of calls in this dataset resulted in an EcoShare acceptance.

[1] Volume of EcoShare Acceptances over Time

From 2017 to 2018, there was a significant decline in the volume of acceptances. From a volume perspective, the month with the most acceptances was in March of 2017 with 1,424 acceptances, which made up over 50% of all the calls for that time period. The following year was when the dip started to sustain, where the proportion of acceptances declined from above 50% to 25% in early 2018. The decline went to as low as 1% in December of 2018. Since 2019, the acceptance rate was still relatively low with occasional peaks such as in Fall of 2019. However, the acceptance rate has not recovered to as high as in the first year and 2 months of this dataset.

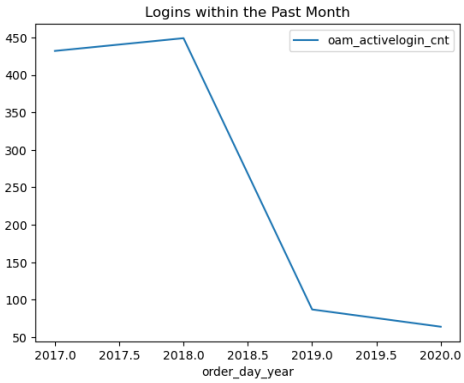


Product Patterns

Volume-wise, most EcoShare offers are accepted for calls for the Reliant Flex and Secure 12 plan. Here are the trends for 5 of the Reliant plans and the acceptance rates by year - more or less, the trends across these plans seem to be quite similar, where there is a steady decline into 2020.

[2] Acceptance Rates for 5 Reliant Product Plans

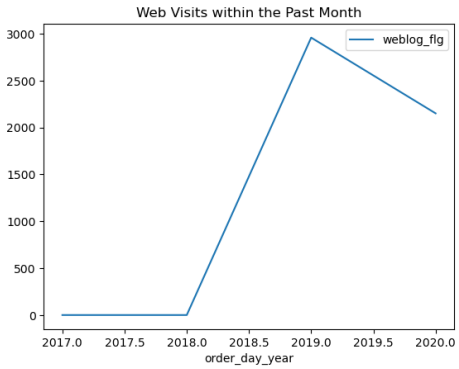
Web Activity



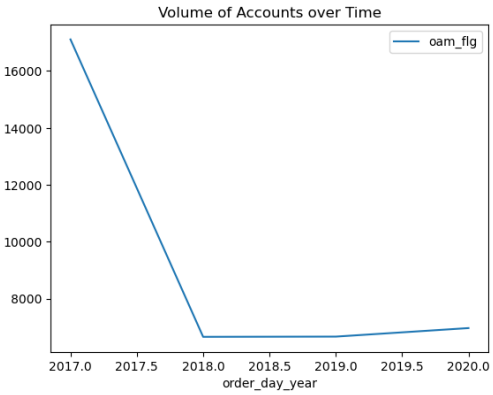
The number of times the customer logged into Reliant's website within the past month seems to follow a similar declining trend as the acceptances but the dip occurs the year after.

[3] Number of Logins Over Time

For whether the customer visited Reliant's website within the past month, the trend here seems to be almost the opposite as compared with acceptances - the peak from 2018 to 2019 corresponds with a decline in acceptances. Maybe there were more issues with the plans and more people were logging into the website, hence less attention on EcoShare. The dip afterwards into 2020 could be related to the pandemic, where there was not as much activity.



[4] Number of Times the Customer Visited the Website within the Past Month over Time



On the left is a distribution of whether the customer has an account or not - it seems to follow a similar trend as acceptances over time initially and may imply a relationship between these two variables:

[5] Number of Calls with Customers that have Accounts over Time

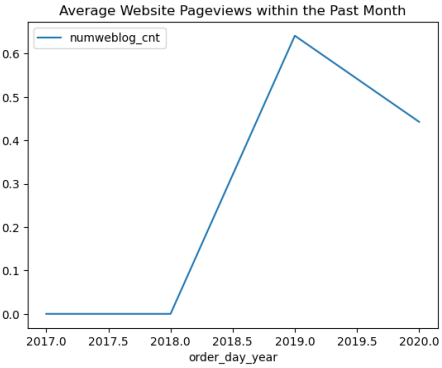


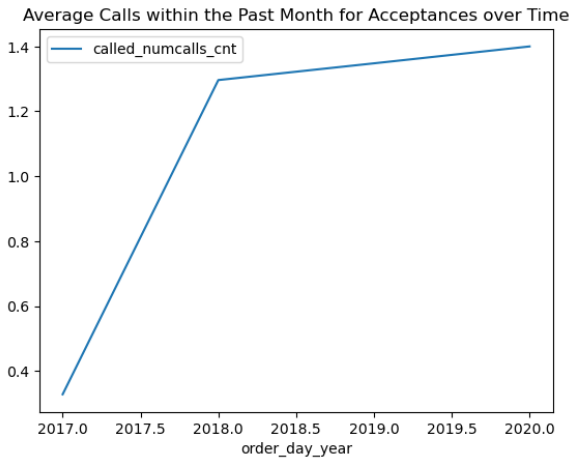
Figure 6 shows the number of pageviews the customer went through on the website - the trend is similar to number of website logins:

[6] Number of Pageviews over Time

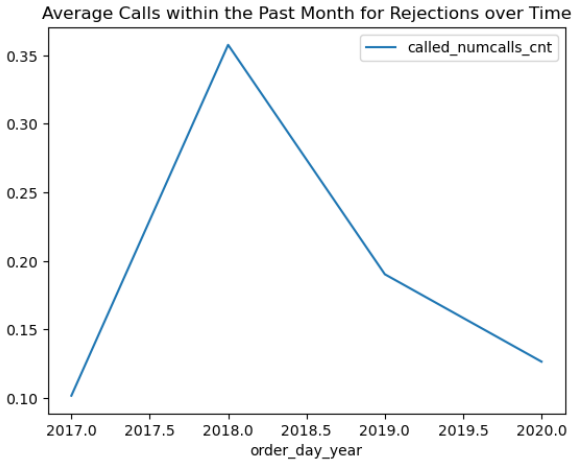
Calls

In terms of the average number of calls within the past month, the chart below on the left shows calls for customers who accepted EcoShare, whereas the right shows the trend for those who didn't. It seems that for those who accepted, the call volume increased between 2017 and 2018, which is when the acceptances decreased from a volume perspective. However, the average number of calls increased within those two years. This might indicate that even though volume decreased, some customers may have called more frequently and they may have been more

likely to accept EcoShare, given potentially more exposure to other programs during those interactions..



[7] Number of Calls for EcoShare Acceptances

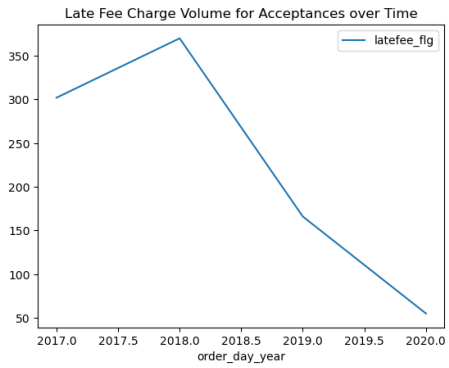


[8] Number of Calls for EcoShare Rejections

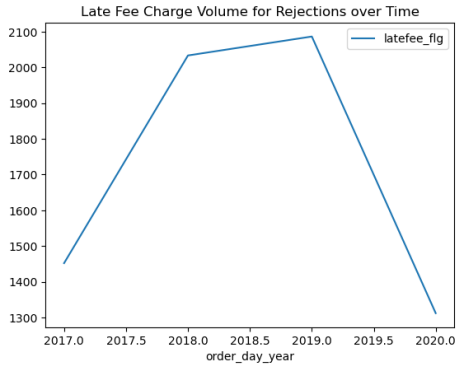
The field, called_flg, which is whether the customer called within the past month, also follows a similar trend for both acceptances and rejections - there may be evidence of multicollinearity for the two call variables.

Payment and Usage

Whether the customer was charged a late fee for not paying their Reliant electricity bill on time within the past month shows an increase from 2017 to 2018, but during that time the acceptances decreased volume wise - possibly an indicator that those who were charged a late fee are not as likely to accept EcoShare:

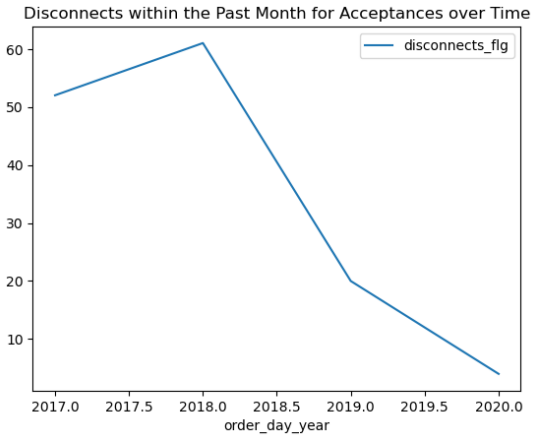


[9] Late Fee for EcoShare Acceptances over Time

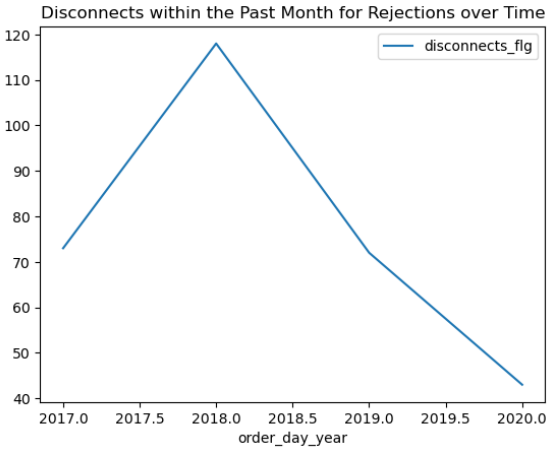


[10] Late Fee for EcoShare Rejections over Time

Whether the customer's electricity service was disconnected within the past month (disconnects_flg) and if the customer received a notice for not paying their bill (disconotice_flg) seem to have high correlation with each other, as the trends over time for EcoShare acceptances and non-acceptances are very similar. In particular, for customers who accepted EcoShare, there is a peak in 2018 but then a decline into 2020.

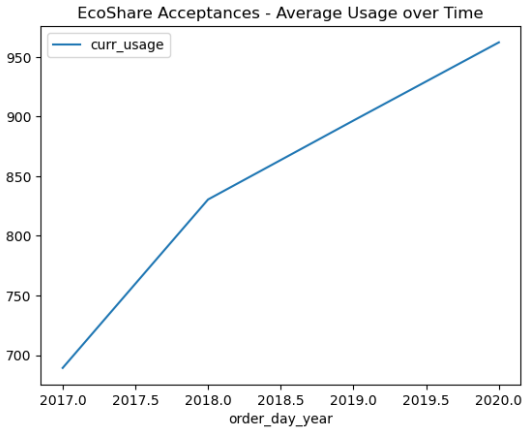


[11] Disconnect Volume for EcoShare Acceptances

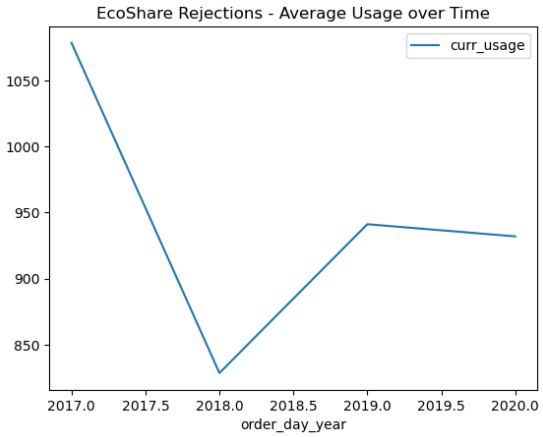


[12] Disconnect Volume for EcoShare Rejections

The usage trend seems to show an opposite trend (see [13] and [14] below) - this may indicate that more usage may mean more engagement with the product and hence knowledge of other programs going on such as EcoShare.

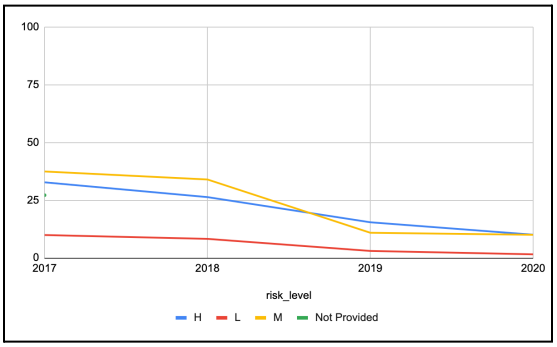


[13] Usage for EcoShare Acceptances over Time



[14] Usage for EcoShare Rejections over Time

Risk Level

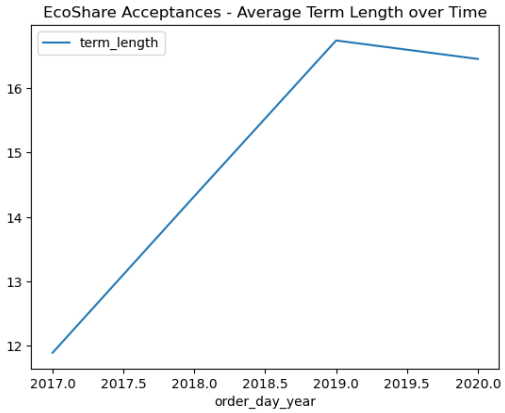


There doesn't seem to be much of an association between EcoShare acceptances and risk level, as the trend of acceptance proportion declined across all levels over the last few years. See the "risk" tab in the file called "eda" in the repository.

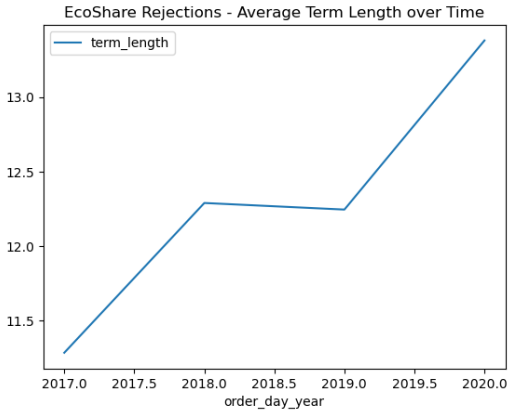
[15] Acceptance Proportion by Risk Level over Time

Term Length

Term length has been a steadily increasing trend over time for both types of customers, but there is a slight dip (see Figure [16]) starting in 2019 for customers who accepted EcoShare:



[16] Average Term Length for Acceptances



[17] Average Term Length for Rejections

Location

40.5% of all calls are with customers in Harris county, followed by about 11% in Dallas county. For the marketing area associated with the customer's service address (dma), Houston is where most customers are located at 52% of all calls, followed by Dallas at 28%. In terms of city, most of the calls with customers are located in Houston at 30%, followed by Dallas at 6.6%. This is expected since the company is located in Houston.

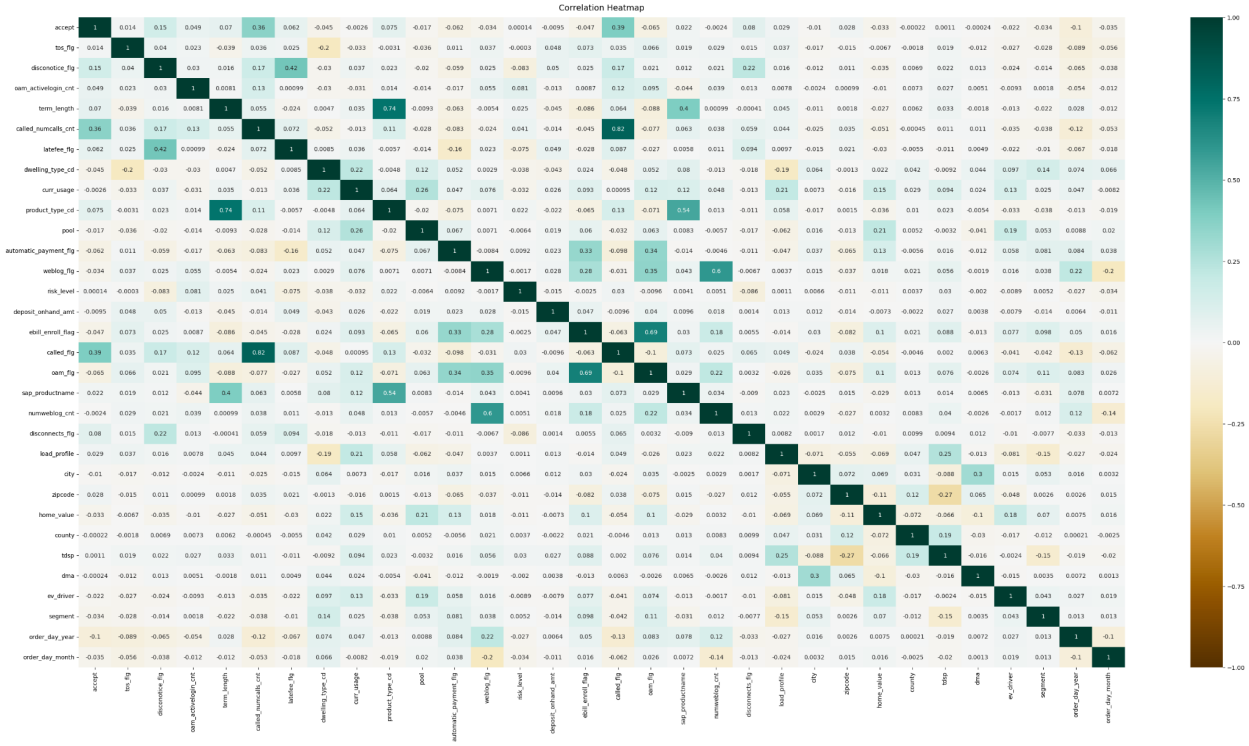
Preparing the Data

These are the changes, imputations, and assumptions for fields that required transformations - most of these variables were converted to categorical variables in order to be in the right format to build the data models on:

- tos_flag: Assumed blanks are “No” and converted to categorical codes
- disconotice_flg: Converted to categorical codes
- term_length: Replaced “C&” and “MM” with blanks and then imputed with the yearly average term length
- latefee_flg: Converted to categorical codes
- dwelling_type_cd: Converted to categorical codes
- curr_usage: Blanks were imputed with the average yearly usage since the volume of nulls was not small - negatives assumed for solar customers
- product_type_cd: Converted to categorical codes
- pool: Assumed blanks meant no pool so imputed these records with “No” and then converted to categorical codes
- automatic_payment_flg: Converted to categorical codes
- weblog_flg: Converted to categorical codes
- risk_level: Categorized blanks as “Not Provided” and then converted to categorical codes
- deposit_onhand_amt: Replaced blanks with 0
- ebill_enroll_flag: Converted to categorical codes
- called_flg: Converted to categorical codes

- oam_flag: Converted to categorical codes
- sap_productname: Categorized blanks as “Not Provided” and then converted to categorical codes
- disconnects_flg: Converted to categorical codes
- load_profile: Categorized blanks as “Not Provided” and then converted to categorical codes
- city: Converted to categorical codes
- zipcode: Converted to categorical codes
- county: Converted to categorical codes
- home_value: Imputed the overall mean for blanks
- tdsp: Converted to categorical codes
- dma: Converted to categorical codes
- ev_driver: Converted to categorical codes

Correlation Heatmap



[18] Heatmap among all Variables

From the heatmap, the number of calls and whether the customer calls is more strongly associated with acceptances. The feature “disconotice_flg” (whether the customer recently received a disconnect notice from Reliant for not paying their electricity bill) also seems to have some relationship with acceptances. However, there is also some evidence of multicollinearity such as for these pairs:

- product_type_cd and term_length (0.74)
- called_flag and called_numcalls_cnt (0.82)
- sap_productname and term_length (0.4)

- disconotice_flg and latefee_flg (0.42)
- oam_flg and ebill_enroll_flag (0.69)
- oam_flg with automatic_payment_flg and weblog_flg (0.34-0.35)
- numweblog_cnt and weblog_flg (0.6)

There is also strong inverse correlation including:

- tos_flag and dwelling_type_cd (-0.2)
- latefee_flg and automatic_payment_flg (-0.16)
- zipcode and tdsp (-0.27)

The fields customer_id, meter_id, and order_day were removed. The training dataset was then trimmed to only have calls from March 2018 and onwards, as the proportion of acceptances was significantly higher only for the first year and early parts of the second year in the dataset. Since then, the acceptances seem to have been on the lower end with rare peaks to above 30% (see "acceptances over time" tab in the sheet in "eda" in "package.zip" within the repository). Because of the sustained decline, I took out the first year and two months of the training set. Also, because there was extreme class imbalance (only about 17% of the calls in the training dataset resulted in EcoShare acceptances), oversampling using the SMOTE (synthetic minority oversampling technique) technique was done to even out the distribution of the classes by replicating examples of the minority class. The methodology that was followed is linked here.

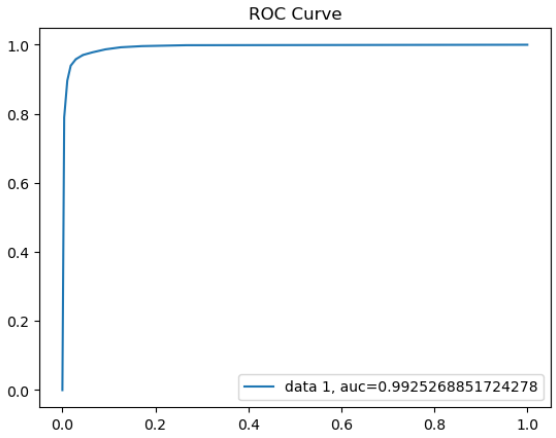
Data Modeling

After feature engineering, four models were built in total and the methods are listed below. For some of the models, feature selection was applied. The training dataset was then split into its own "training" and "test" sets, and then the model was built based on the training set. Cross-validation was applied and then the evaluation metrics including accuracy, AUC, and precision and recall were generated.

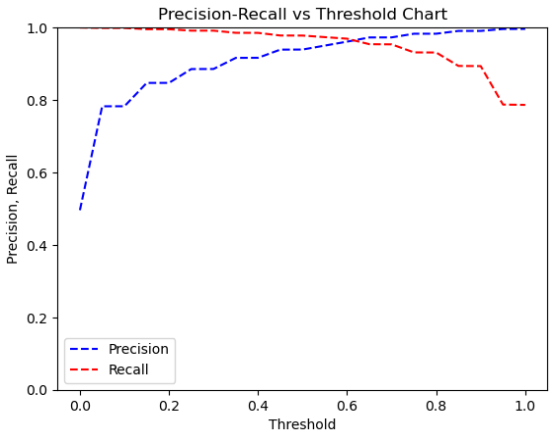
Model 1: Extra Trees

The first method that was built was the Extra Trees classifier. This is similar to a regular random forest where both classifications are made up of decision trees and then the growing method is the same where at each node, there is random selection of the feature to be added. However, extra trees is based on the original sample of data rather than on bootstrapping. The selection of where to split is selected randomly.

The model leveraged 10 estimators, and the criterion that was used was entropy, which is a measure of disorder in the dataset**.** Cross validation was also applied to see the generalized accuracy score for other datasets. The accuracy for cross validation was around 96%. Based on the test dataset, the accuracy was close to 95-96% and the AUC was also around 0.96. Below shows the precision-recall chart.



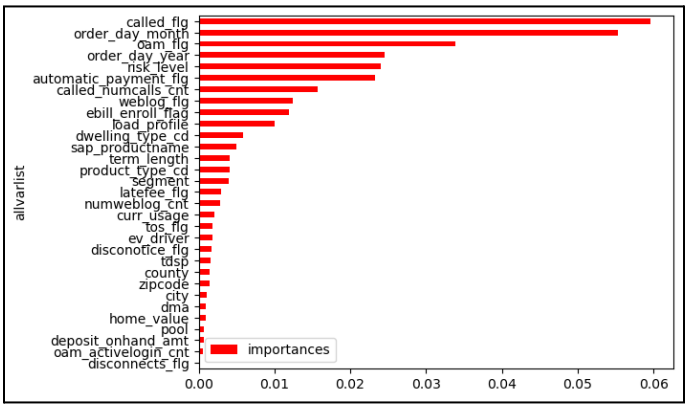
[19] ROC Curve for Model 1



[20] Precision-Recall Chart for Model 1

Model 2: Extra Trees with Top Features

The next method that was built was the same model but included just the top features. Here is the importance plot that was generated based on the model:



[21] Importance Plot for Model 2

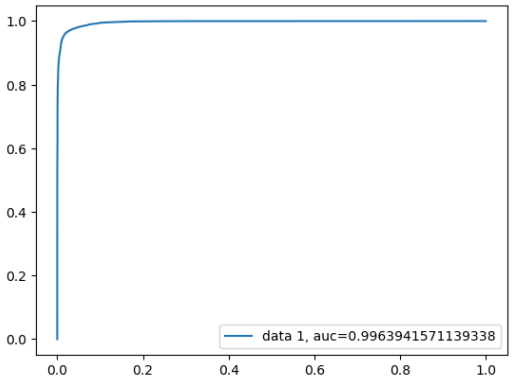
It seems like whether the customer called within the last month, the month and year of the order date, oam_flg, automatic_payment_flg, and the risk level (whether the customer is at risk of not paying their electricity bill, grouped into low risk (L), medium risk (M), and high risk (H)) are some of the top features that impact the acceptance. Whether the customer had a call with Reliant seems to be quite indicative of acceptances; possibly as more calls come in, the customer may be more up to date with the programs at NRG including EcoShare. We use the top eight features as part of the model. Again, the training dataset was split 80/20 and then the model was built and was also cross-validated. The resulting testing accuracy was around 92% with an F1 score of around there as well. So far, calls, payment issues, and product type seem to consistently be the top features that are strongly associated with EcoShare acceptance.

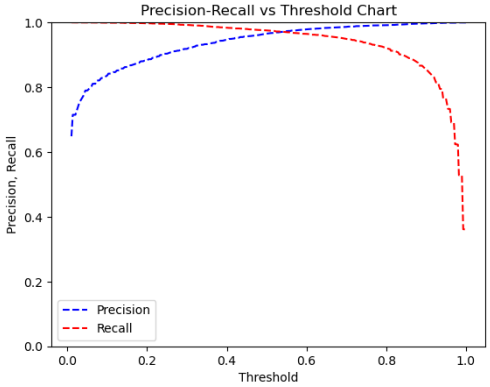
Model 3: Random Forest with All Features

Here, a random forest model was built, incorporating all features. The cross-validation accuracy was around 97% and the precision and recall were around 97% as well. The difference with Extra Trees is that the Random Forest model is built based on bootstrapped sets of data. The

AUC is also quite high at almost 0.97.

The ROC curve is shown below along with the precision and recall chart:



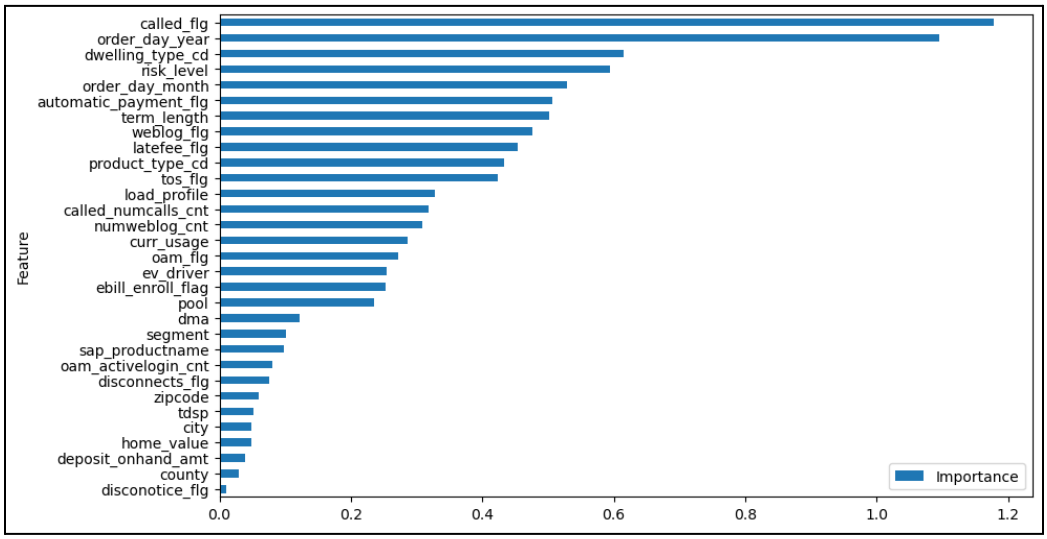


[22] ROC Curve for Model 3

[23] Precision-Recall Chart for Model 3

Model 4: Logistic Regression with Top Features

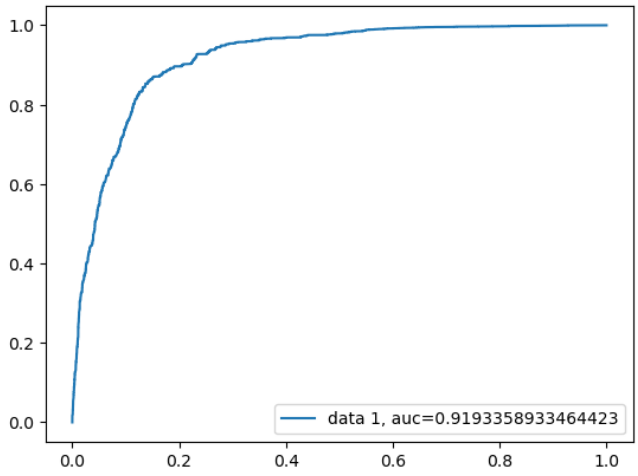
Lastly, a logistic regression model was built for the top features. To prevent overfitting, only the more important or strongly associated features were selected. Below shows a graph of the importance of each feature (scaled), which was calculated based on the absolute values of the coefficients of the model output:



[24] Importance Graph for Method 4

The eight features that were used were: called_flg, order_day_year, risk_level, order_day_month, automatic_payment_flg, weblog_flg, dwelling_type_cd, and term_length. The data was also scaled before being built on along with k-fold cross-validation. The average accuracy was around 0.83, and the F1 score was close to 0.86.

Below shows the AUC chart:



[25] AUC Chart for Model 4

Results and Discussion

We evaluate the model performance by mainly the overall accuracy of the split test set within the training data, AUC score, and the F1 score, which combines precision and recall. A high accuracy does not mean a good AUC score nor a good model, and so it's essential to also compare the performance across models by using AUC and precision and recall. "Good" AUC scores also vary by industry (article reference here). In this case, we also try to optimize both precision and recall. However, we may prioritize a higher rate of positive classifications so as not to lose out on opportunities of participants in the EcoShare program that would benefit both financially and for the environment.

Below are some of the metrics across all models:

```
Method 1 Accuracy: 0.9576301875860072
Method 2 Accuracy: 0.9241688998334179
Method 3 Accuracy: 0.9699427826464837
Method 4 Accuracy: 0.8568117621496343
```

```
Method 1 F1: 0.9581335432620053
Method 2 F1: 0.9269211977385357
Method 3 F1: 0.9704605309986475
Method 4 F1: 0.8581473774843941
```

```
Method 1 Precision: 0.941036080303243
Method 2 Precision: 0.8903729172176673
Method 3 Precision: 0.9665988963113564
Method 4 Precision: 0.8631888317413666
```

```
Method 1 AUC: 0.9577967737534226
Method 2 AUC: 0.924020543134325
Method 3 AUC: 0.9698859216752301
Method 4 AUC: 0.9193358933464423
```

[26] Comparison across all Models

Based on the methods leveraged above, the third model, which is the random forest method, performs the best across all metrics above. This model was used to predict for the test set provided by NRG. Please see the probabilities of EcoShare acceptance in the file called "deborah_chang_predictions.pkl" in the repository here. The test set was cleaned the same way as the training dataset before it was run through the model.

It's clear that overall, calls seem to have the strongest association with acceptance. When more calls are made by a customer, they may be more engaged and up to date with the latest plans and products. Even though they may have called for a separate purpose such as for billing or payment issues, when pitched with an offer more times, that may have increased the likelihood of accepting EcoShare.

The retention call center data uncovered many insights on the customer, their buying history, and preferences that may lead to acceptances. More insights can be gleaned for additional purposes including improving customer service, ROI, and more.

Conclusion

Overall, in predicting whether a Reliant customer would accept EcoShare, the random forest model performed the best in terms of accuracy, AUC, and precision and recall. Call history, product type, and payment issues show stronger relationships with EcoShare acceptance. Addressing the class imbalance also significantly helped with modeling the data.

The practicum provided good real-life experience into taking a dataset and figuring out how to analyze and model it using external context and assumptions. Future work could be to dive more into other products and see what customers usually prefer, or the models could be refined for overfitting and regularization. Other models could also be looked into, and there could be more exploration on the features to see if more can be pulled and investigated, including text fields where sentiment analysis can be leveraged. Overall, the practicum provided ample experience in working with real-life datasets, navigating unexpected challenges with modeling, and learning how to communicate findings and insights.

Appendix

Data Dictionary

Field Name	Description
order_day	The date when the customer was offered EcoShare in the retention call center
accept	Whether the customer bought EcoShare during the call
tos_flg	Whether the customer transferred their service with Reliant Energy from one home to another prior to the call
disconotice_flg	Whether the customer recently received a disconnect notice from Reliant for not paying their electricity bill
oam_activelogin_cnt	The number of times the customer logged into Reliant's website within the past month
term_length	The length of the electricity contract they purchased from Reliant
called_numcalls_cnt	The number of times the customer called Reliant's call center within the past month
latefee_flg	Whether the customer was charged a late fee for not paying their Reliant electricity bill on time within the past month
dwelling_type_cd	Whether the customer lives in a single family home (S) or multi-family home (M)
curr_usage	The amount of electricity consumed at the customer's home during the previous billing cycle
product_type_cd	Whether the customer is enrolled on an electricity contract (TERM) or not (MTM)
pool	Whether the customer's home has a swimming pool
automatic_payment_flg	Whether the customer is enrolled on Automatic Payment
weblog_flg	Whether the customer visited Reliant's website within the past month
risk_level	Whether the customer is at risk of not paying their electricity bill, grouped into low risk (L), medium risk (M), and high risk (H)
deposit_onhand_amt	The deposit amount the customer paid when they first enrolled with Reliant Energy for electricity service
ebill_enroll_flag	Whether the customer receives their electricity bills electronically
called_flg	Whether the customer called Reliant within the past month
oam_flg	Whether the customer has an online account on Reliant's website
sap_productname	The name of the electricity plan the customer is enrolled on
numweblog_cnt	The number of pageviews the customer experienced on Reliant's website within the past month
disconnects_flg	Whether the customer's electricity service was disconnected

	within the past month
load_profile	A code that represents the shape of the customer's electricity consumption pattern. R1 means the customer has gas heating, R2 means the customer has electric heating, and other load profile values are rare.
city	The city associated with the customer's service address.
zipcode	The zipcode associated with the customer's service address.
home_value	The estimated value of the customer's home.
county	The county associated with the customer's service address.
tdsp	The utility that serves the customer's home.
dma	The marketing area associated with the customer's service address.
ev_driver	Whether the customer is estimated to drive an electric vehicle.
segment	NRG's custom, propriety segmentation of its customers. Each segment value represents a unique type of consumer in the Texas market.
customer_id	A randomly generated number that represents the account holder who was offered EcoShare on the call.
meter_id	A randomly generated number that represents the meter(s) of an account holder who was offered EcoShare on the call.

Works Cited

Administration, Call Center. "How Can Data Analytics Help You Identify and Address the Root Causes of Call Center Attrition?" *How Data Analytics Can Reduce Call Center Attrition*, www.linkedin.com, 23 Apr. 2023, www.linkedin.com/advice/0/how-can-data-analytics-help-you-identify.

Brownlee, Jason. "How to Develop an Extra Trees Ensemble with Python." *MachineLearningMastery.Com*, 26 Apr. 2021, machinelearningmastery.com/extra-trees-ensemble-with-python/.

"How to Analyze Call Center Data to Improve Efficiency." *RSS*, www.invoca.com/blog/how-to-analyze-call-center-data-to-improve-efficiency. Accessed 12 Nov. 2023.

"ML: Handling Imbalanced Data with Smote and near Miss Algorithm in Python." *GeeksforGeeks*, GeeksforGeeks, 11 Jan. 2023, www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/.

NRG Energy, Inc. "NRG Energy, Inc. to Acquire Vivint Smart Home, Inc.." *NRG Energy*, www.nrg.com/about/newsroom/2022/41771.html. Accessed 12 Nov. 2023.

NRG Energy, Inc. "Welcome to NRG." *NRG Energy*, www.nrg.com/. Accessed 12 Nov. 2023.

Zach. "What Is Considered a Good AUC Score?" *Statology*, 9 Sept. 2021, www.statology.org/what-is-a-good-auc-score/.