Midterm Progress Report

Fayaaz Khatri
10/05/2023

Overview

- Project context
- Project objective
- Current progress
  - Categorical variables
  - Feature Engineering
  - Random Forest Classifier
  - Cross validation and parameter tuning
- Planned work and next steps
  - Updates and recut of training data
  - Consider time-series nature of dataset and objective
  - Iterate on model and results

Project Context

reliant®
an NRG company

- Reliant - an NRG brand that serves electricity in Texas
- Reliant interested in recommending additional products to existing customers in smart, targeted ways
- One such product is EcoShare - a carbon offset program offered for a small monthly fee
- Provided with call center training dataset for existing Reliant customers
- Call center agents encouraged to upsell EcoShare to eligible customers
- 94,601 call records
- 18,191 calls led to enrollment ~ 19% success rate

Project Objective

- Develop model that predicts likelihood of EcoShare enrollment, given that an eligible Reliant customer calls into the call center
- Will be used in NRG's Cross Serve Personalization Engine to help call center agents intelligently upsell additional products
- Deliverables
    - Probability of EcoShare enrollment for each call record in test set
    - Packaged and documented code to implement model in production

Current Progress - Data Preprocessing

Categorical Variables

- Converted into dummy variables
- The field `sap_productname` has 800+ levels - still looking into ways to cluster these
- Several fields required imputing missing values

Feature Engineering

- Parsed date into several numeric date parts
  - Month
  - Year
  - Day of year
- Geocoded zip codes to numeric latitude and longitude

Current Progress - Classification Model

Random Forest Classifier

- Best performing model so far with 84% accuracy with a random train-test split on training data
- Top 5 in feature importance
  - Day of Year
  - Longitude
  - Latitude
  - Year
  - Risk Level of L

Cross Validation and Parameter Tuning

- Marginal to no improvement when tuning the following RF parameters
  - Max depth of tree
  - # of features eligible for best split
  - # of trees
- Mean CV test scores ~ 83%

Planned Work and Next Steps

Updates and Recut of Training Data

- NRG sponsor Sean mentioned some data quality issues around customer account and billing account granularity in training data
- Dataset will be recut based on date as well, to avoid potential “data leakage” - exposure to future data to infer on past data
- Expecting an updated training dataset in early October

Planned Work and Next Steps

Time-series Nature of Dataset and Objective

- Random Forest Classifier train-test split was random, not cut based on date, so possible data leakage and inflated accuracy results
- Redo train-test split to have training records earlier than test records - consider backtesting
- This is better aligned with NRG's objective in production, since only historical call data will be available to model and predict a new incoming call

Planned Work and Next Steps

Iterate on Model and Results

- Experiment with other classification models using train-test split based on date
- Look into insights from other scoring methods like AUC, ROC, precision, recall, etc
- Consider dimensionality reduction and variable selection - most features have low importance
- Look into additional feature engineering, particularly based on industry knowledge
- Finalize and package code with documentation