Wayfair Challenge

Using Data to Bring Customers Home

by Kyle Maxwell

Kyle Maxwell

About Me

Hello!

My name is Kyle Maxwell. I'm a fourth year undergraduate student studying Software Engineering at California Polytechnic State University. I heard about this challenge in our weekly computer science email newsletter. I've had a ton of fun working with this dataset and participating in the challenge!

OVERVIEW
THE PROCESS

Data Exploration

See WayfairExploration.ipynb
Worked on cleaning the data, dealing with missing values, handling outliers, normalizing, and performing feature selection.

Experiments

Experimented with different methods for feature selection and different types of models for classification and regression.

Validation

See WayfairDevelopment.ipynb
Selected the final model, data preprocessing pipeline, and tuned hyperparameters based on a generated training-validation split.

Prediction

See WayfairChallenge.ipynb
Trained models on all of the training data and generated final predictions

EXPLORATION AND EXPERIMENTATION

Beyond tediously playing around with handling outliers and missing values my main challenge was feature selection.

Originally I was performing feature selection via forwards stepwise regression using BIC and R^2 as metrics. I found that was slower and no better than just training a preliminary gradient boosting model to determine the most import features.

Another aspect I had to deal with was the skew of the revenue distribution. With an enormous skew of 25.6, I decided to correct for it by applying log(1+x) to the revenue. As such, I applied the inverse upon inference.

MODEL SELECTION AND VALIDATION

Regarding selecting the models, I played around with various techniques. I found that for the classification of conversion, gradient boosting performed the best. For predicting the revenue, I found that a gradient boosting model worked just as well as a fully connected neural network, but was much faster to train.

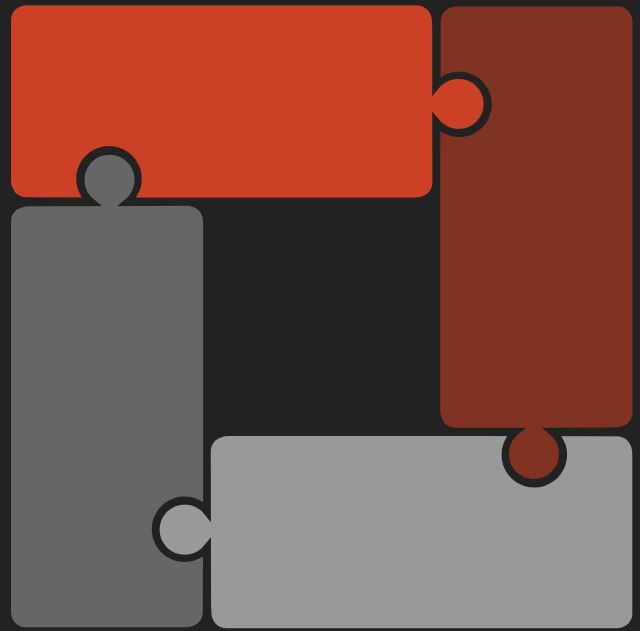
I created a 3:1 training to validation split to optimize the structure and hyperparameters. For the classification of customer conversion I was observing precision, recall, and their F1 harmonic mean. For the prediction of revenue I observed RMSE.

PREDICTION
PUTTING IT TOGETHER

In the end, here's how the process looks.
First, I preprocess the training data and perform feature selection for the conversion classification problem. Then, I train the conversion model.

Next, I take the subset of the ground truth samples that did convert, and perform feature selection for the predicting revenue problem. Then, I train the revenue model.

For testing, I similarly preprocess the testing data and predict which samples converted. Finally, I perform the revenue prediction on that subset of samples predicted to convert.



THANK YOU!

Thanks for taking the time to host this challenge! Let me know if you need any clarifications or have any questions. I'm looking forward to hearing the results of the competition.

I can be best reached at:
kylemaxwell7@gmail.com