Predicting recipe ratings

Abstract

A vegetarian recipe creator is curious to find out if ratings on his newly developed recipes could be predicted. As a data scientist, my goal was to use linear regression to predict recipe ratings by scraping vegetarian recipe data from allrecipes.com. Data scraping was performed using selenium and beautiful soup and imported into a pandas dataframe. Data was cleaned and exploratory analysis was conducted on jupyter notebooks. Finally, cleaned data was modeled using linear regression and polynomial regression.

Design

Data was scraped using selenium to obtain 1000 URLs and beautiful soup to scrape specific field values for each recipe by iterating over the URLs. Data cleaning consisted of addressing nulls and outliers. Categorical features were dropped and numerical features were standardized. Data was split into 80/20 for train and test sets respectively. Baseline linear regression, linear regression with transformed target variable(rating) and polynomial regression with lasso regularization were conducted along with cross validation to come up with a robust modeling technique.

Data

1000 recipes were scraped with 11 features related to a recipe such as title, rating, prep time, cook time, recipe category, yield, servings, calories, protein, carbohydrates, fat, and sodium. Most of the features were continuous numerical variables. Categorical variables were discarded for modeling. Rows with nulls for recipe rating were discarded and nulls in nutrient info were replaced with average values from the respective subcategories.

Algorithms

Feature Engineering

1. Correlation matrix and VIFs of most of the features showed no relation to the target and among other features except for calories

2. Calories showed strong collinearity with other features and so was dropped from modeling
3. Standard scalar was used to standardize all the features
4. Target variable (rating) displayed no linear relationship with other features and was heavily skewed to the left. Boxcox transformation was used to normalize the data

Models

Logistic regression, polynomial regression with Lasso regularization modelling techniques was opted for modeling. Unfortunately, the target variable had no linear relationship with any of the features and hence linear regression performed poorly. Polynomial regression performed well on training data only indicating that the model was overfit. Hence, Lasso regularization was performed to address the overfit but could not be eliminated. Kfold cross validation was conducted between the models and linear regression performed better than polynomial.

```
Linear Regression scores:       [0.014, 0.012, -0.04, 0.013, 0.01]
Polynomial Regression with Lasso scores:        [0.02, -0.017, -0.021, 0.0, -0.024]

Linear Regression cv r^2: 0.002 +- 0.021
Polynomial Regression with Lasso mean cv r^2: -0.008 +- 0.016
```

But it is important to note that none of the modelling techniques would be recommended to predict the ratings because of the low accuracy. However, future goals of the project include gathering more diverse data and features that can explain the nature of ratings on the recipes.

Tools

- Selenium and Beautifulsoup for scraping
- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Tableau for visualizations

Communication

Presentation to the group outlining some key findings.