BrainStation Capstone

Liana Principe

4/8/2023

**Oh SNAP! Curbing America’s Hunger**

**Intro and Objective**

With the help of modern farming methodologies, technology, and machines, food production is at an all time high in human history. Global food production is currently able to produce enough food to feed the entire population with a 30% surplus. And yet, 34 million people spanning across every county in the United States do not have access to a sufficient amount of food to meet basic needs. Several studies have been conducted to explore varying demographic factors and the relationships they might have with food insecurity. In 2012, the ERS and USDA administered an extensive survey to American households across the country to collect comprehensive data about household demographics, relationship to food, and purchasing behavior. The National Household Food Acquisition and Purchase Survey (FoodAPS) covered content such as household composition, employment and income, expenses, access to food, and participation in the Supplementary Nutrition Assistance Program (SNAP), formerly known as food stamps. Using SNAP as a proxy, it then becomes possible to link household socio-economic variable data to food insecurity. With a robust governmental system such as SNAP, a processing system and model that labels applicants as eligible or not eligible for the provided benefits already exists. Rather than researching the socio-economic features that affects whether a household is likely to become food insecure, this research deviates slightly by identifying which features impact the likelihood of a household to actually utilize the resources if they are experiencing food insecurity. This research question is devised under the presumption that not all who are eligible for SNAP benefits will utilize these resources, and the target group of interest becomes those who ultimately will utilize the resources available.

The purpose of this research is to use demographic and socio-economic variables to create a machine learning model that is capable of identifying households that are likely to utilize SNAP resources. This research statement is under the guise that those who utilize SNAP resources are eligible and are thus considered to be experiencing some level of food insecurity. A high performing model will be able to identify which features are most impactful on determining if a household will or will not utilize SNAP resources, which can be integral in alleviating or assisting those in need. Harnessing this knowledge opens up opportunities for community outreach through events, informational campaigns about resources, or can even instigate policy decisions.

The underlying process of this study is to train a model to classify a household as “has used SNAP resources” versus “has not used SNAP resources”, with the intent of using this to identify households who are likely to utilize these same resources in the future. A public version of the survey responses that has been cleaned and organized is available online that contains the records of all 4,826 households. Although the FoodAPS provides extensive data covering numerous topics, only the socio-economic features from the survey were used to train the model in efforts to preserve reproducibility and mimic data collection that is easily obtainable by standard cross documentational practices. Any self-reflective responses or personal surveys are omitted from the model building process.

**Data Cleaning and Preprocessing**

After taking the consolidated public version of FoodAPS and removing rows with null values, the dataset still contained 4798 rows of data with 96 socio-economic variables. Preliminary feature selection was performed through a Decision Tree process to identify any features that were reported to have no significant value in classifying households who utilize SNAP resources. Those features were removed from the dataset and this process was repeated until all features were reported to have a significance that was at least greater than 0, indicating that feature had some value in classifying the target output.

From here, the data was further transformed by scaling and principal component analysis (PCA) to experiment how feature engineering would affect the model. By running small, preliminary logistic regressions, it was found that the original dataset and scaled data produced similar model scores, but PCA had the lowest scores across several evaluation scores. Thus, PCA was not considered when moving forward in the model design process.

**Model Evaluation**

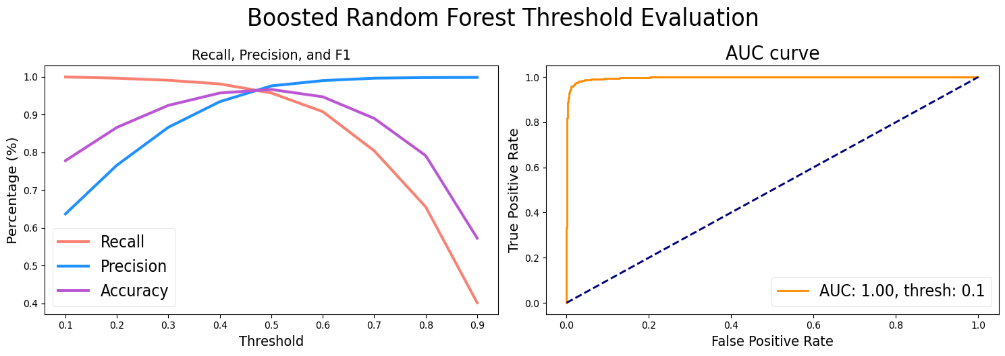
While designing and choosing certain models, several scoring metrics were taken into consideration to have well-rounded and accurate results. The metrics most recorded in the study were Recall, Precision, Accuracy, F1, ROC\_AUC, and a confusion matrix. Given the main objective and purpose of this study, recall was selected as the primary metric of focus as it measures the ability of the model to properly identify all households in the dataset that utilized SNAP resources. In other words, out of all the households in the dataset that are recorded as utilizing SNAP resources, how many can the model correctly identify. In this case, it is preferred to overestimate, meaning the model classifies more households as utilizing the benefits than there really is, as opposed to underestimating, meaning the model fails to identify some households that were in need of SNAP resources.

**Model Building**

Once equipped with a dataset of significant socio-economical variables and a primary evaluation metric, the next step is to design the model. As the objective of this research is to identify households that are likely to utilize SNAP resources, classification models were implemented in the model building phase. Logistic Regression and Support Vector Machine (SVM) models were trained in the case that the data is linear and a boosted Random Forrest model was trained in the case of non-linear data. For each of these three models, a hyperparameter grid search was conducted to identify the model and its underlying parameters that would best be able to classify SNAP households from the training data. The grid search simulated hundreds of parametric combinations for each model, applying cross validation, and scoring the results according to the specified metric. The outputs provided better context to different model combinations which was integral to choosing a model that would best be able to classify households that utilize SNAP resources.

**Threshold Tampering**

In addition to model building and optimizing, the last step used in this study was threshold tampering. Changing the threshold value of the model impacts the way the model makes the final classification, meaning that altering the threshold will cause the model to be more or less conservative with its final classification output depending on the change to the threshold. The threshold value was taken into consideration in this study to improve the performance of certain metrics of the model. The figure below shows an example of the threshold tampering process. Specified metrics are plotted for each threshold value, giving a better understanding at what threshold value the performance metrics are optimized at. With threshold tampering, the model is fully optimized, complete, and ready for final model evaluations and final tests.

****

**Results and Conclusion**

In the final test run, both models had a recall over 80% with the logistic model being slightly higher than the boosted random forest by 2%. The general metric scores are comparable between the two models, however the logistic regression model was able to classify slightly more households that utilized SNAP resources than the boosted random forest model. The figures below show the final model evaluations in addition to the top 10 features the model deemed as significant when running the classification. The graphs show that there is some overlap between the top 10 features, although the majority of them are different.

As the model stands, both models are able to correctly classify 82-84% of households that have utilized SNAP resources. Future implications of this research would take this model and use it to identify households in current society that may be in need of assistance in alleviating food insecurity. This can take form with community outreach programs, awareness campaigns, or simply resource informational tactics.

A recall above 80% is sufficient on paper, although that indicates that 20% of the households who might benefit from SNAP resources would go unclassified. The exploration of alternative or more mature models might increase performance results, as would a deeper investigation to the features in the model. While there is room for growth in this study, the work conducted in this research provides good foundational support on the topic of alleviating food insecurity.

