FOOD INSECURITY

GROUP 6

WOYRAM ALMARANTE

BRENTON FENDER

MICHAEL CICCHINO

EMERY SCOTT

VASILIY KONDRATYEV

Topic and Reasoning

- Food insecurity is defined as “a lack of consistent access to enough food for every person in a household to live an active, healthy life”. (USDA, 2023).
- Based on the increasing rate of inflation following the recent global events, how secure are households in the USA?
- What are the factors that impact food insecurity?
- Is there enough data to reveal trends and predict how access to food sources impact food insecurity.

Research Question

Main Hypothesis:
How does physical access to food through restaurants and grocery stores impact food insecurity?

Sub Hypotheses:
- How is fast-food expenditure related to food insecurity?
- Does restaurant availability contribute to lowering food insecurity?
- Which has a higher positive impact on food access: having more farmers markets or more general grocery stores?
- How does agritourism influence food insecurity?
- How do community supported farms versus regular farms have an impact on food insecurity?

Data Exploration and Processing

- The agricultural theme steered the team to the USDA website regarding food insecurity. We discovered the Food Environment Atlas.

- Our original data was transformed and cleaned using Python and Pandas within Jupyter Notebook. This included conversion of data types, removal null and duplicate values, and formatting.

- Physical locations became the key features of our analysis.

- The team refined the feature variables to restaurants and stores to create a concise dataset for the machine learning model. Local was dropped due to incomplete data.

Description of the Data

2020 Food Environment Atlas from United States Department of Agriculture includes varying data from 2007-2019. Additional food insecurity data has been to added to make total representation from 2009-2019.

Stores:

- Groceries
- Supercenters and club stores
- Convenience stores
- Specialized food stores

Restaurants:

- Fast-food
- Full-service

Agriculture:

- Community supported agriculture
- Farms with direct sales
- Farmers markets

Data Limitations

- Data skewed towards reductions in food insecurity.
- Variable dates not corresponding with food insecurity data.
- Food insecurity data only reported at state level with variables reported in more localized areas.

Description of Analysis Phase

1. The team chose to predict the % change in Household food insecurity from 2014-16 to 2017-19.

2. Team 6 recognized a skewed /imbalanced nature of the y-variable. For this timeframe food insecurity only increased for 5 states CT, MD, NV, SD and WV.

3. The team utilized Balanced Random Forest Classifier to account for the imbalance as well as selected a % change to predict movements in Food Insecurity as opposed to a Binary increase or decrease.

Model Overview

- Target:

  Household food insecurity (change %),2014-16 to 2017-19*

- Features:

  Grocery stores/1,000 pop (% change), 2011-16

  Supercenters & club stores/1,000 pop (% change), 2011-16

  Convenience stores/1,000 pop (% change), 2011-16

  Specialized food stores/1,000 pop (% change), 2011-16

  Fast-food restaurants/1,000 pop (% change), 2011-16

  Full-service restaurants/1,000 pop (% change), 2011-16

Model - Confusion Matrix & Feature Importance

Confusion Matrix

- Reviewing the confusion Matrix to the right we can see the accuracy of the model's predictions. Specifically, we see that it has issues calculating a the median change of food insecurity. Smaller and larger changes we predicted more accurately.

	Predicted 0	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Actual 0	59	0	0	0	0
Actual 1	0	98	0	0	0
Actual 2	26	10	159	8	6
Actual 3	4	2	0	140	1
Actual 4	6	5	2	1	144

Feature Importance

- Reviewing the feature importance to the right. Our team reached a counterintuitive conclusion. It appears the presence of Convenience Stores per 1,000 people % Change from 2011 to 2016 is the most important feature in predicting food Insecurity % Change from 2014-2016 to 2017-2019.

```
# List the features sorted in descending order by feature importance

sorted(zip(rf_model.feature_importances_, X.columns), reverse=True)
```

```
[(0.19032823788844175, 'Convenience stores/1,000 pop (% change), 2011-16'),
 (0.18782585436779842,
  'Full-service restaurants/1,000 pop (% change), 2011-16'),
 (0.18570094023090214, 'Fast-food restaurants/1,000 pop (% change), 2011-16'),
 (0.17765527742884302, 'Grocery stores/1,000 pop (% change), 2011-16'),
 (0.13042692111945925,
  'Specialized food stores/1,000 pop (% change), 2011-16'),
 (0.12806276896455548,
  'Supercenters & club stores/1,000 pop (% change), 2011-16')]
```

Model – Methodology & Results

- The team selected to use a Ensemble Learner to analyze the dataset. Specifically, we chose to use a Balanced Random Forest Classifier (BRFC) to account for overfitting.

- The Team utilized sklearn's Balance Accuracy Score. We can see the model's has a high average recall obtained for each class.

```python
# Calculated the balanced accuracy score
from sklearn.metrics import balanced_accuracy_score
y_pred = rf_model.predict(X_test)
balanced_accuracy_score(y_test, y_pred)
```

```
0.9249077815366956
```

- When calculating out imbalance classification report we can see the precision, recall and F1 scores that the model does a reasonable job at predicting movements in food insecurity. The model's accuracy decreases with more drastic changes. (% Changes > 4%)

```python
# Print the imbalanced classification report
from imblearn.metrics import classification_report_imbalanced
print(classification_report_imbalanced(y_test, y_pred))
```

	pre	rec	spe	f1	geo	iba	sup
-4	0.62	1.00	0.94	0.77	0.97	0.95	59
-3	0.85	1.00	0.97	0.92	0.99	0.97	98
-2	0.99	0.76	1.00	0.86	0.87	0.74	209
-1	0.94	0.95	0.98	0.95	0.97	0.93	147
0	0.95	0.91	0.99	0.93	0.95	0.89	158
avg / total	0.92	0.89	0.98	0.90	0.94	0.87	671

Dashboard

Created in Tableau

Final Thoughts and Conclusions

- Two main takeaways:
  - The per capita percentage change for Convenience Stores is the **best metric** to predict Food Insecurity changes on a state level compared to other features, including grocery stores, restaurants, fast food and etc.
  - Geographic location challenged with food insecurity have a higher reliance on convenience stores as a source of food.
- Looking back: anything the team would have done differently?
  - Handling missing data
    - Due to time constraints, the team opted to delete NAs, which resulted in a 30% dataset reduction. With more time, we can potentially “keep” more data and improve our model using imputation or interpolation.
  - Obtain more granular datasets
    - Our datasets had state-level food insecurity information. We believe that running our analysis on county-level data could benefit our research.
- Recommendations to future researchers…
  - Try to replicate our findings/results using deep learning model

The Technologies, Languages, Tools And Algorithms That The Team Used Throughout The Project

- Collaboration tools
  - Slack
  - Zoom
  - GitHub
- Data exploration and analysis
  - MS Excel
  - Jupyter Notebook
  - Python libraries:
    - scikit-learn
    - pandas
    - imbalanced-learn
- Database
  - ERD Generator - QuickDBD
  - PostgreSQL
  - pgAdmin
- Visualisation and presentation:
  - Tableau
  - MS PowerPoint and Google Docs

REFERENCES

- USDA, 2023

  https://www.ers.usda.gov/webdocs/publications/45020/30967_err141.pdf
- USDA ERS - Data Access and Documentation Downloads

  https://www.ers.usda.gov/webdocs/publications/44906/6893_err125_2_.pdf?v=5244
- 2009-2011 Food Insecurity Data

  https://www.ers.usda.gov/webdocs/publications/45020/30967_err141.pdf
- 2011-2013 Food Insecurity Data
  https://www.ers.usda.gov/webdocs/publications/45265/48787_err173.pdf
- 2014-2016 Food Insecurity Data

  https://www.ers.usda.gov/webdocs/publications/84973/err-237.pdf?v=219.4