PURDUE UNIVERSITY | The Data Mine

Web App Implementation with Logistic Regression

Reis McMillan, Sara Xiao, Ananya Srivastava, and Parth Ranade

Keystone COOPERATIVE™

INTRODUCTION AND BACKGROUND

Keystone Cooperative, formerly Co-Alliance Cooperative and Ceres Solutions Cooperative, is the seventh largest agricultural cooperative in the United States. Such a size, however, is not without its challenges. Many of the Keystone customers have multiple accounts through the cooperative with no indication that such accounts belong to the same customer. To maximize the potential of the account data that Keystone has, our team was tasked with developing a web app to help business users identify accounts belonging to the same customers.

Throughout the development of the project, the team had the following goals in mind:

- Ensure the web app had a simple, user-friendly interface
- Identify key features of data which indicated potential account matches
- Develop a reliable method for predicting account matches

To achieve the above goals the team received the following data to work with:

- Individual account data
  - Customer name, address, longitude and latitude, associated Keystone location
- Invoice split data
  - Many invoices are billed to multiple accounts
- Existing account groups
  - Some groups of accounts have already been self-identified by Keystone customers

CONCLUSIONS

Having developed the logistic model, the team was able to make a few observations:

- Distance between accounts was *NOT* a significant feature of data
- Address fuzzy matching scores were a more reliable indicator of whether an account pair was a match than name matching scores
- The logistic model favored false negatives over false positives
  - The percentiles for false positives and false negatives:

False Positives	
Percentile	Probability
0th	50.41%
25th	55.27%
50th	64.37%
75th	76.07%
100th	96.57%

False Negatives	
Percentile	Probability
0th	51.77%
25th	69.67%
50th	80.26%
75th	90.47%
100th	99.69%

  - The percentiles for true matches:

True Positives	
Percentile	Probability
0th	50.53%
25th	96.88%
50th	99.78%
75th	99.98%
100th	100.00%

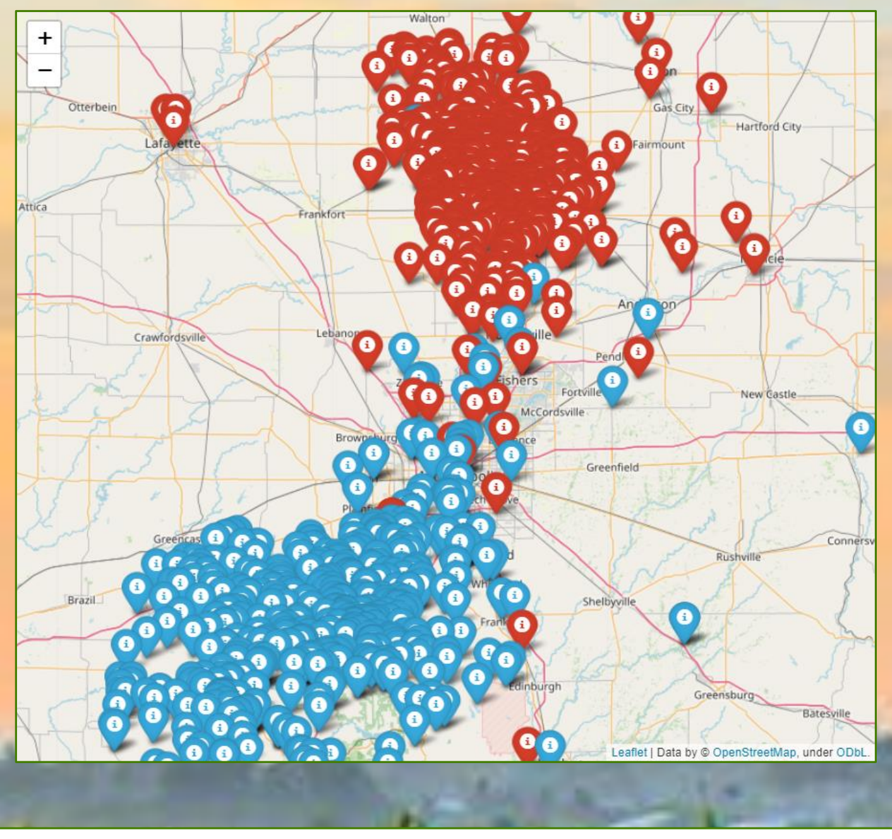
True Negatives	
Percentile	Probability
0th	51.71%
25th	81.66%
50th	89.61%
75th	93.85%
100th	99.62%

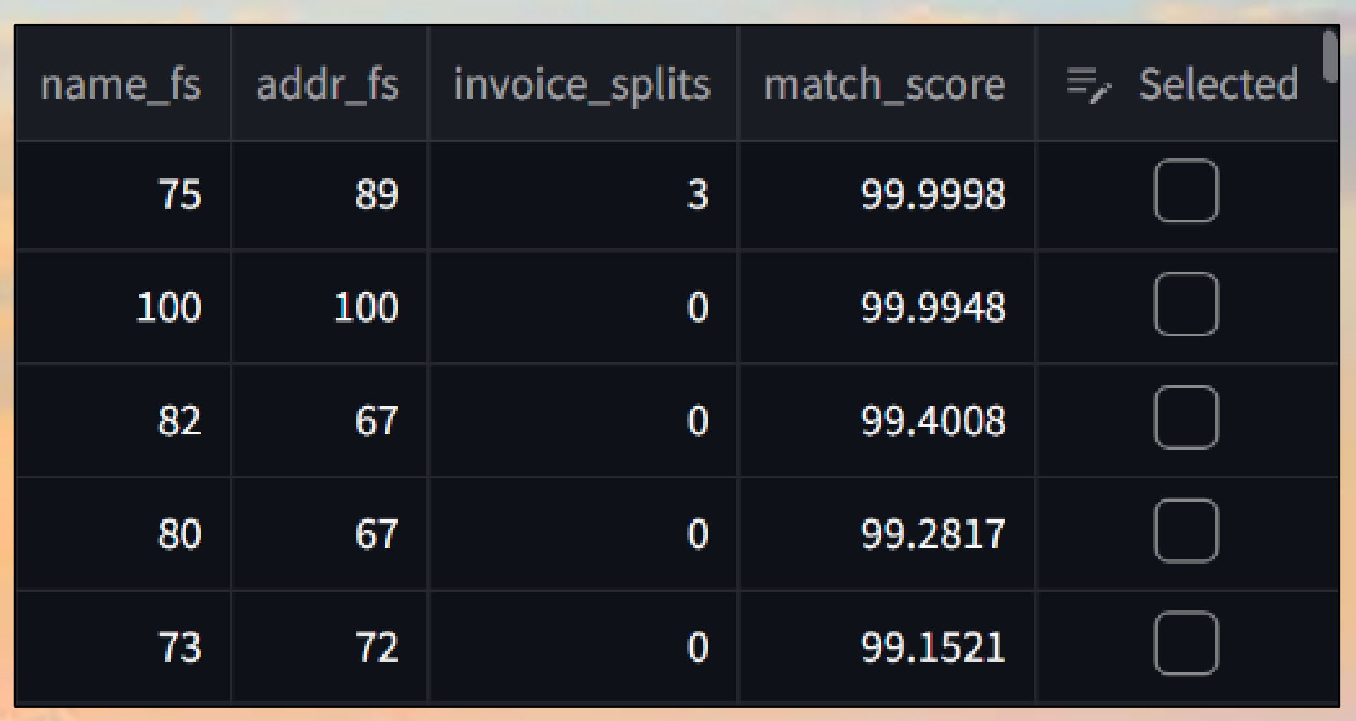
  - The ideal cutoff for recommending matches in the web app should be 85% probability of a true match
- Invoice splits as a feature of data should be treated as numerical data and not categorical data

After the team's final iteration of developing the logistic model, the team was able to achieve 91.3% accuracy using a 75-25 test-train split.

WEB APP DESIGN

The team opted to use Streamlit, a Python framework for building data science and machine learning web applications, for its ease of use and ability to deliver high level functionality within the web app. The team wished to include an interactive map within the web app; originally PyDeck, a mapping library, was selected due to its ability to integrate with Streamlit. However, it had limited user interactivity, so the team opted to use Folium, which could also be deployed in a Streamlit environment.





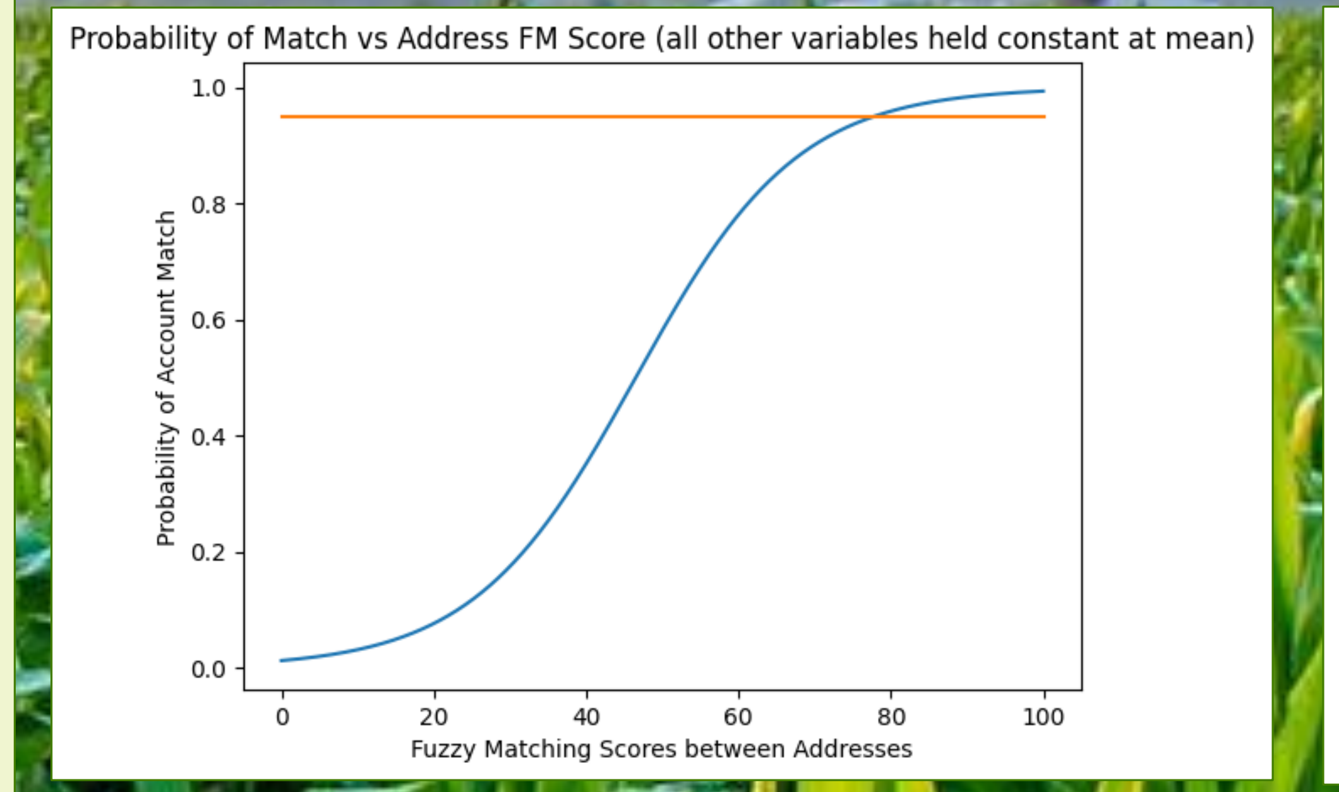
name_fs	addr_fs	invoice_splits	match_score	Selected
75	89	3	99.9998	□
100	100	0	99.9948	□
82	67	0	99.4008	□
80	67	0	99.2817	□
73	72	0	99.1521	□

LOGISTIC MODEL DEVELOPMENT

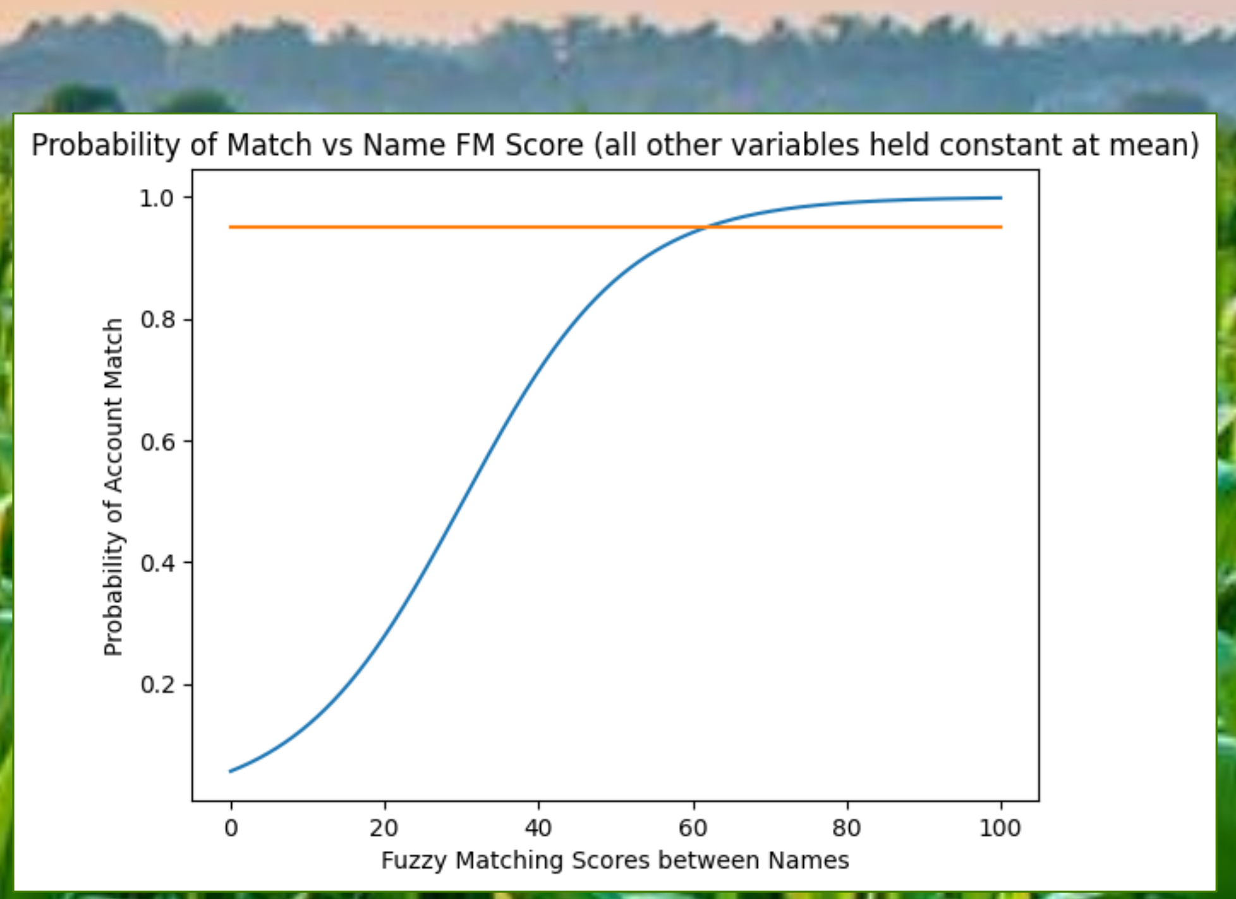
The goal of the web app is to assist business users with identifying accounts which belong to the same customer. Because of the availability of existing linked account data, the team opted to use logistic regression to create a model which could recommend potential account matches to business users.

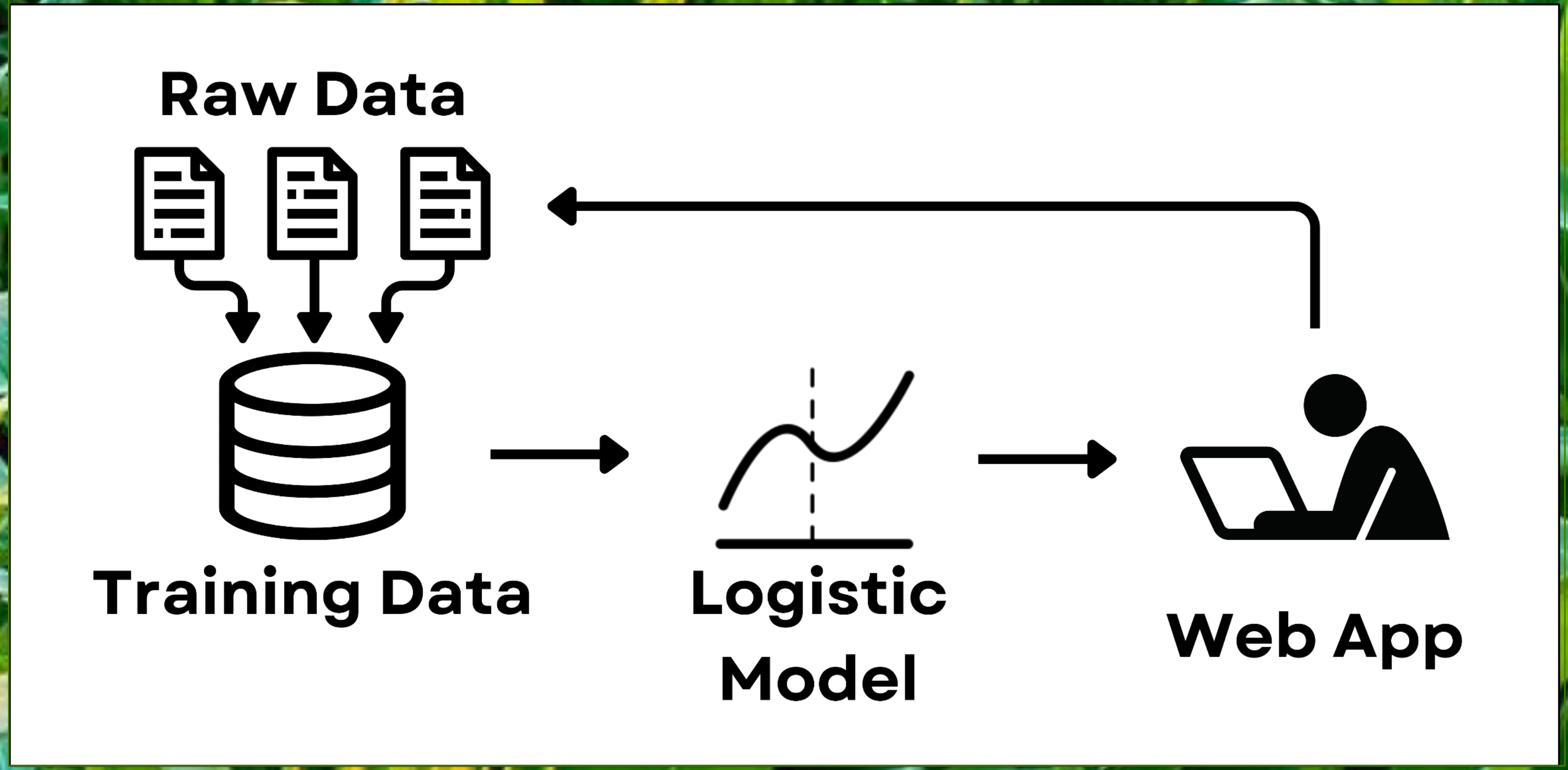
The team pulled and merged data from three different data sets to create a training and test data set for the model. Each row of the resulting data set represents an account pair which is either a match or non-match. Key features of data include fuzzy matching (text comparison) scores (FM score) between account names and account addresses, the distance between accounts, and the number of invoices which two accounts have shared.

Probability of Match vs Address FM Score (all other variables held constant at mean)

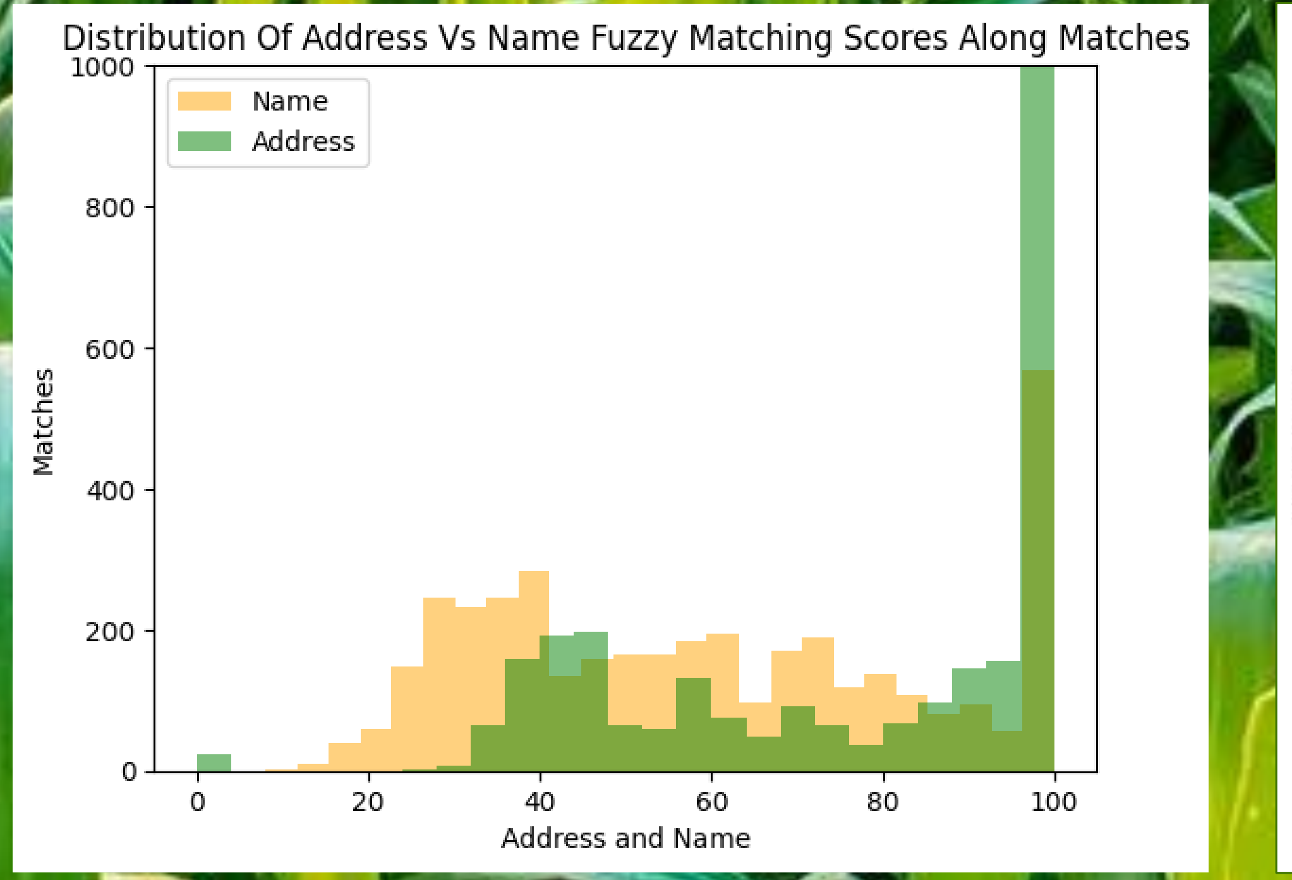


Probability of Match vs Name FM Score (all other variables held constant at mean)

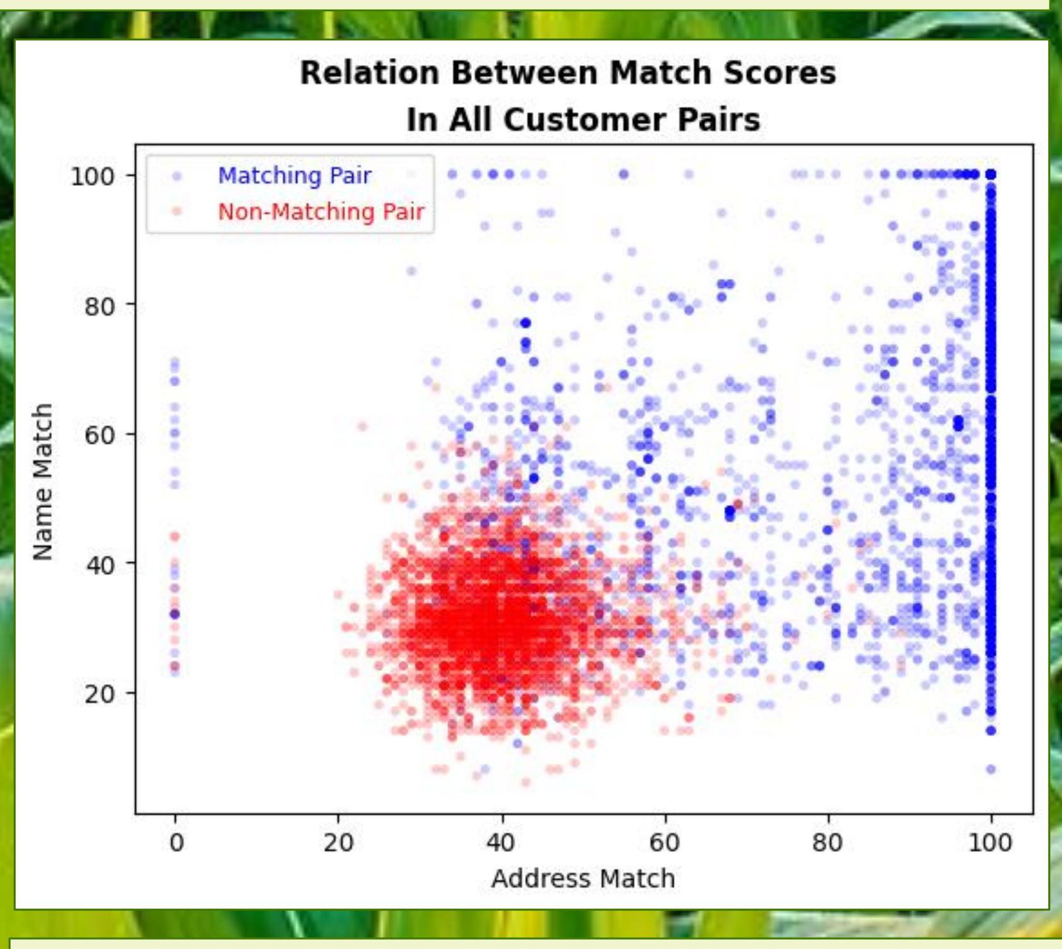




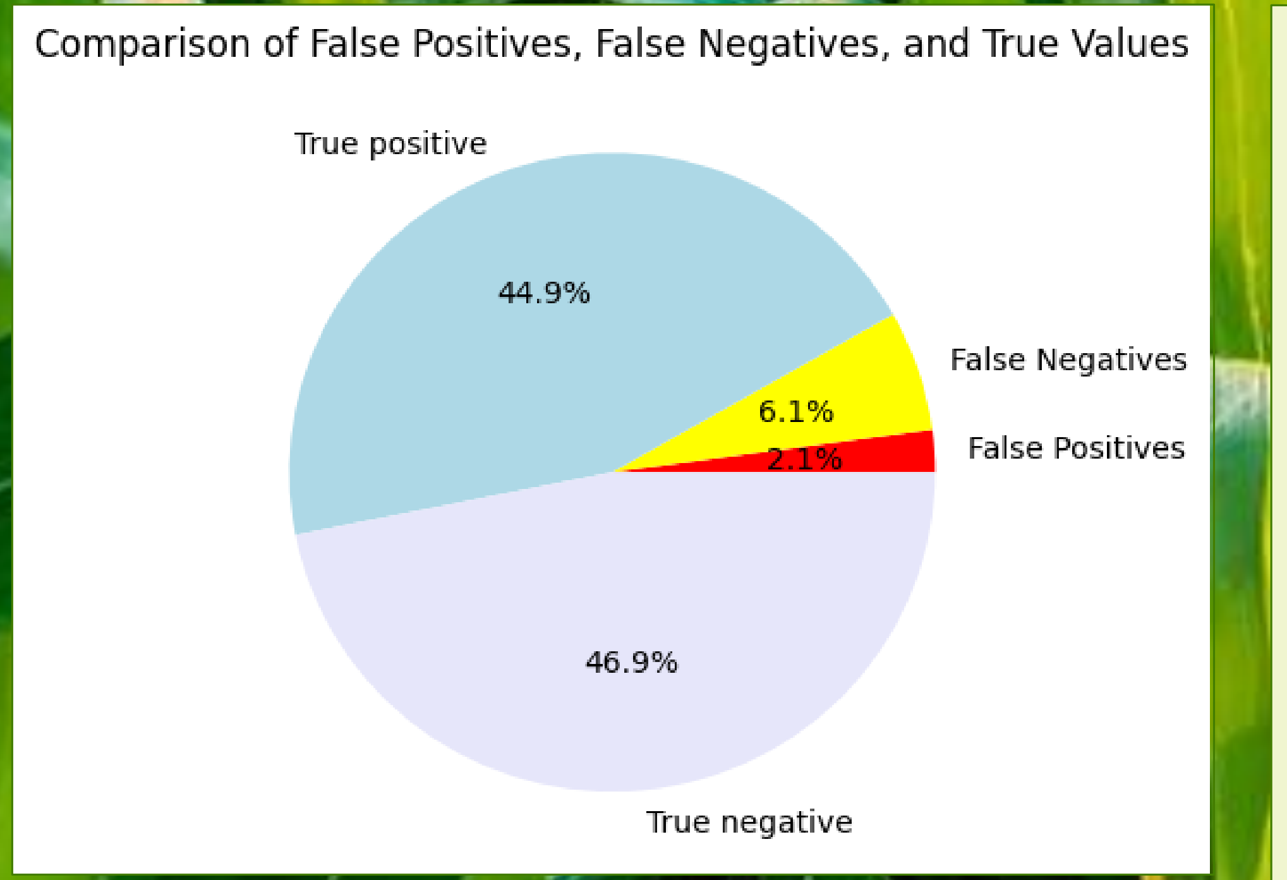
Distribution Of Address Vs Name Fuzzy Matching Scores Along Matches



Relation Between Match Scores In All Customer Pairs



Comparison of False Positives, False Negatives, and True Values



FUTURE GOALS

The result of the teams work this year is a web app which is ready to be deployed into a production environment. By choosing to use Streamlit, the web app can deploy seamlessly in Keystone's Snowflake cloud-computing and data platform. Such integration will allow the web app to work in real-time with Keystone Cooperative's databases.

To begin such an integration into the Snowflake environment, the team will need to make considerations regarding multiple users and ways to optimize the logistic model further. Also, the team will need to familiarize itself with Snowflakes various APIs, which are essential to integrating the web app into a Snowflake environment.

ACKNOWLEDGEMENTS

The team would like to thank Tyler Netherly, our corporate mentor with Keystone for all the help he was able to give the team throughout the project. We would also like to thank Tracy Soper and Jeff Griffeth at Keystone for allowing such a project to be possible.

Much of this year's work would not have been possible without the assistance of The Data Mine staff like Cai Chen and Doug Crabill. Thank you and all other TDM staff who work hard to support this corporate partnership!

The Data Mine Corporate Partners Symposium 2024