Identifying Anomalous Transactions

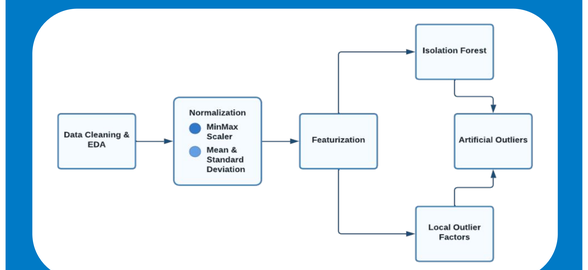
Brian Che, Dingan Jiang, Mira Patel, Ruoxin Wang, Justin Vo
Sponsor: Ari Polakof
Mentor: Erika McPhillips

UCSB Data Science Capstone 2023
University of California, Santa Barbara
&
Appfolio, Inc.

UC **SANTA BARBARA** | Data Science Initiative

Overview

- **Appfolio, Inc.** is a company that offers industry-specific business software solutions, services, and data insights to the real estate markets with the mission of revolutionizing business
- The **goal** of this project is to **detect anomalous transactions with outlier detection models**
- By implementing and comparing various algorithms, we were able to come up with the final explainable analysis method for further usage in the Appfolio Property Manager software



Data

- Data comes from Appfolio records of 3 property management companies
- Corresponds to **payables** and **receivables** that each property manager has for all their properties.
- The variables in this data include the transaction **amount**, transaction **date**, transaction **description**, **property name**, **property address**, **vendor name**, and **General Ledger (GL) Account**
  - The GL Account is a category assigned to the transaction (i.e. water, electricity, etc.)
- **Unsupervised**: No labels for what is/isn't an outlier

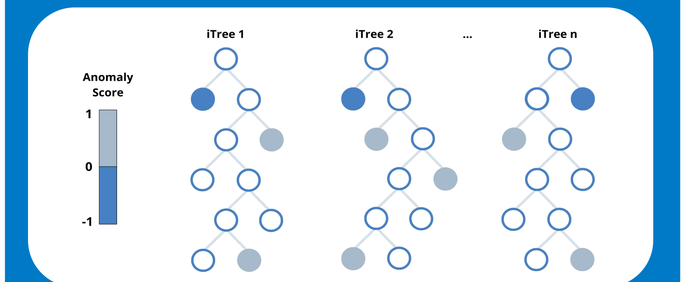
Methodology 1: Normalization

- The range of the amount column was widespread so we normalized the data so it was all on the same scale
- Performed **standard normalization** with respect to unique groupings of property names and GL accounts
- Allows us to look at **local outliers** with respect to groups and not just global outliers
- Some data points would become **NaN** when normalizing with standard deviation equal to 0 (**can't divide by 0**)
  - Only **one transaction** in the grouping
  - All transactions in the grouping have the **same amount value**

Methodology 2: Featurization

- We engineered **new features** that gave us more information about individual transactions based on the given data
- The **mean**, **median**, and **standard deviation** of different groupings:
  - Property name and GL Account
  - Property name, GL Account, and Vendor Name
  - Property Management Company and Vendor Name
  - Vendor Name
- Whether the transaction was **recurring** (monthly, quarterly, etc.)
- Whether the transaction was a **utility** based on a list of key terms
- **Difference** of the amount from the **mean and median** of the property name and GL account grouping

Methodology 3: Isolation Forest Modeling



- Anomalies, being 'few and distinct', are particularly susceptible to isolation. In a data-induced random tree, the process of instance partitioning is repeated recursively until all instances are effectively isolated
- Since anomalies are less frequent, they lead to a smaller number of partitions and, consequently, shorter paths in the tree structure. Early in the partitioning process, instances with distinguishing attribute values are more likely to be separated, resulting in noticeably shorter paths for anomalies
- Therefore, when a forest of random trees produces reduced path lengths for specific sites, there is a high likelihood that these sites are anomalies

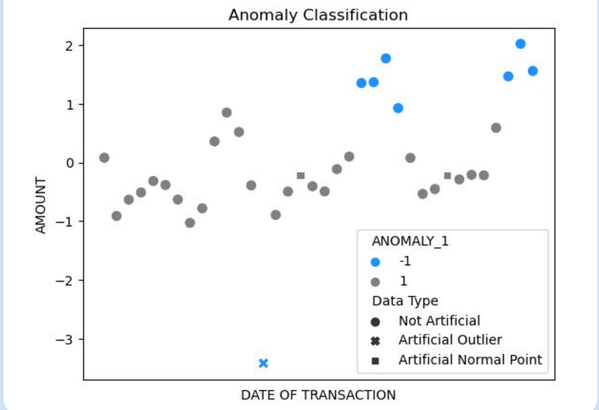
Methodology 4: Artificial Outliers

- We decided to **inject artificial outliers** using a subset of 10,000 observations to **evaluate** our model's ability to **detect anomalies**
- The outliers were generated by **replacing 'AMOUNT' values randomly** outside a specified range determined by the **minimum value**, **lower IQR bound**, **upper IQR bound**, and **maximum value**
- We generated **normal points** using the **median of groupings** and **noise** to assess how the model **distinguishes outliers from typical instances**

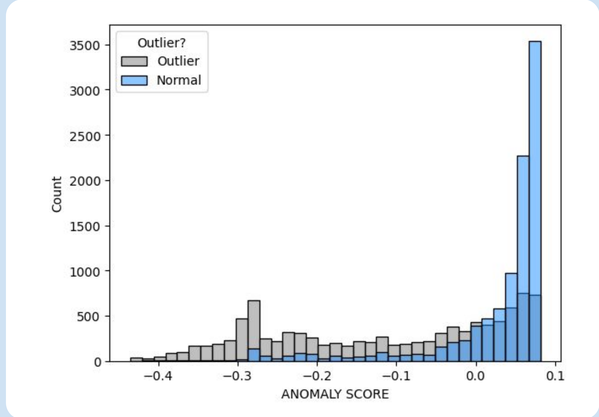
Results

Precision:	75%
Recall:	**73%**
Accuracy:	74%

- After testing various models using the artificial data, we found that our initial model was very effective at detecting extreme/global outliers, but less so at detecting local outliers (within groups).
- We found that a much simpler method, using **the normalized amount column (and difference from median) performed comparably well** and was able to detect moderate/local outliers.
- Given that results are highly dependent on how our artificial labels were generated we can **visually evaluate model performance.**



Artificial outliers (denoted by X) were flagged as anomalies (identified in blue), while artificial normal points (denoted by a square) were not flagged.



The isolation forest assigned anomaly scores to artificial observations, with negative scores corresponding to anomalous data. Here, we can see that most of the negative scores were assigned to artificial outliers and most of the positive scores were assigned to artificial normal data.

Conclusions

Normalization proved to be the most important step for a few reasons:

1. It standardized the scale of the 'AMOUNT' column making comparison between groups possible
2. It best **captured the disparity between a given transaction amount and similar transactions.** Non-anomalous transactions would have values close to zero while outliers would take on more extreme values
3. Finer groupings during the normalization process improved results, likely by prioritizing comparison amongst similar transactions

Featurizing data did not provide marginal benefit since **all observations within a group will have the same features.** The normalized transaction amount serves as the most distinguishing feature and provides the majority of relevant information.

Future Work

Future exploration of this open-ended question may involve:

1. Further exploration of different kinds of unsupervised machine learning methods to detect anomalies
2. Different strategies to evaluate model performance. This may include different ways of generating artificial data or artificial labels
3. Other methods of normalization that preserve information within groups
4. Further consider how training with different number of features would affect the actual performance of the model
5. Explore other possible features that could be engineered to capture information about individual transactions rather than the grouping that the transaction belongs to

References and Acknowledgments

[1] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.

Special thanks to our sponsors, Appfolio, Inc., represented by Ari Polakof and Soeren Thust, for their invaluable support and to our mentor, Erika McPhillips, for her assistance.