Constraint Detection for Outlier Explanation

Shruti Parab ✧ Deepti Chavan ✧ Sushmita Sinha

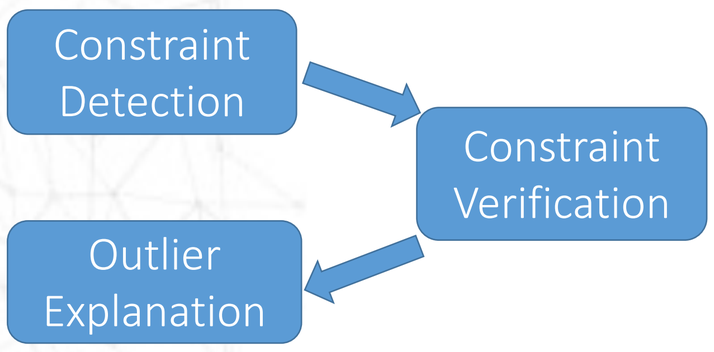
University at Buffalo
The State University of New York

Motivation

- ❑ Constraints govern the data and help find outliers
- ❑ Outlier detection is crucial to ensure data integrity
- ❑ Automate the process of finding constraints and further define trends, make predictions, and uncover root causes for behavior of the data

Introduction

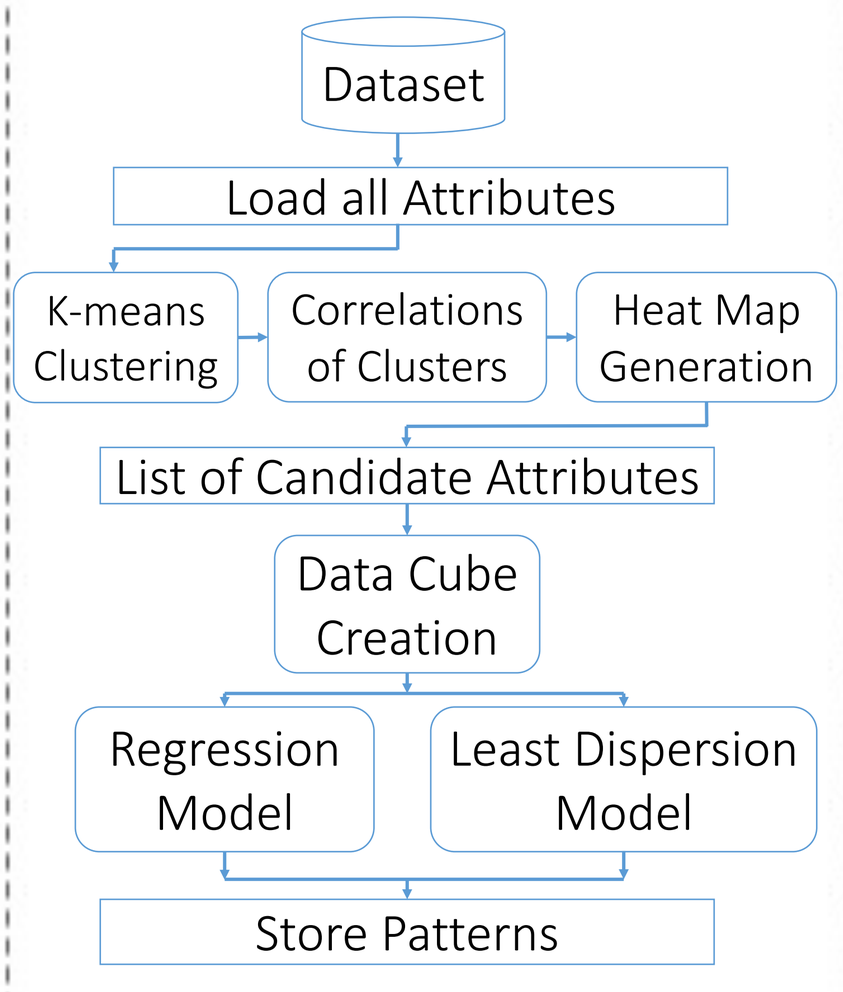
- ❑ The high level aim is to automate the process of finding constraints in the dataset and to optimize the same.
- ❑ Using these constraints, find possible reasons identifying and justifying the presence of an outlier.
- ❑ To determine the statistically significant relationship we use Least Dispersion Method (LDM) to find constant pattern and Linear Regression (LR) for increasing and decreasing pattern



Terminologies

- ❑ Constraint: fixed, variable, aggregate_value, model, metric
- ❑ Pattern: Model that fits on variable vs aggregate values for groups formed by fixed attributes
- ❑ Model: LDM finds the % of standard deviation w.r.t. mean
- ❑ LR fits a regression line with R^2 coefficient as the goodness of fit measure
- ❑ Local Patterns: Patterns discovered for the groups formed by fixed attributes
- ❑ Global Patterns: A summary of how many percentage of groups hold valid patterns

Approach



Optimization

- ❑ Correlation Analysis
  - Find correlation coefficient between pairs of attributes
  - Prune unwanted attributes using pre-defined threshold
  - Explore patterns in reduced space
- ❑ Clustering
  - Form clusters of data points and find correlation coefficient on the cluster points
- ❑ Data Cube
  - Precompute repeated aggregates in the form of a data cube

Results

- ❑ On Loan Dataset with 2M rows

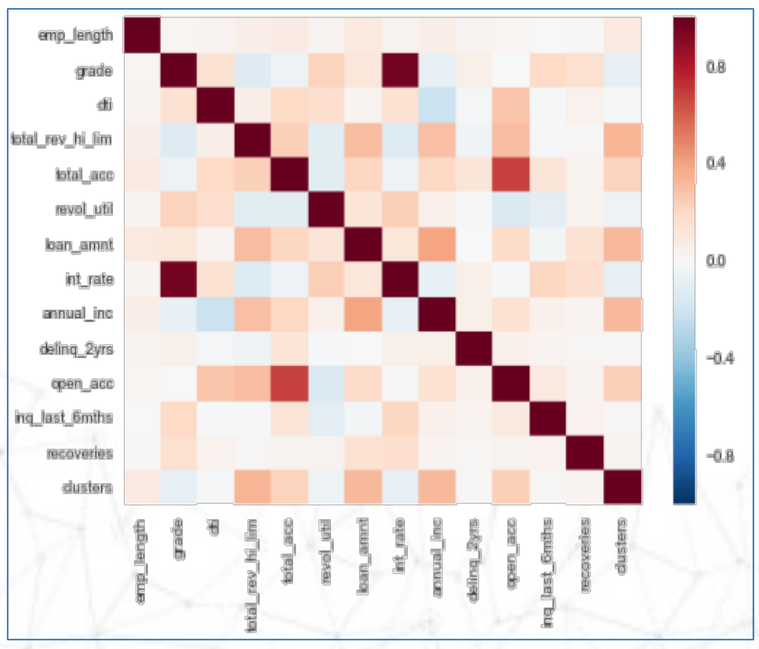
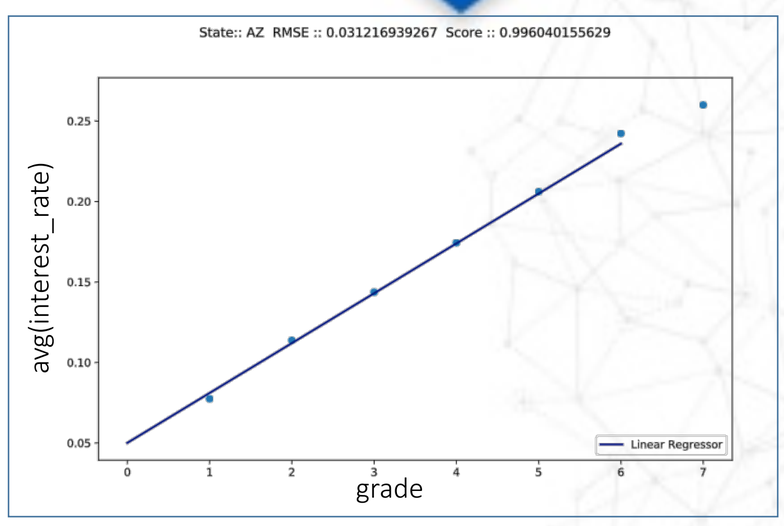


Fig. Heat Map



LR: (f:state, v:grade, agg:interest_rate)

- ❑ Local Pattern: (AZ, grade, avg(interest_rate), 'increasing', 0.99)
- ❑ Global Pattern: (state, grade, avg(interest_rate), 100%)

Key Findings

- ❑ Along with finding correlated pairs from the database our approach further tries to fit a model over those, quantifying it by a goodness of fit score
- ❑ Clustered data points helps in identifying stronger correlated columns, which are otherwise lost in direct matrix correlation
- ❑ With an initial over-head of building the data-cube, pre-computed aggregate values significantly help to improve performance while fitting various models

Future Work

- ❑ Enhancing the readability and usability of patterns discovered
- ❑ Finding Nonlinear Regression patterns

Contributors & References

- ❑ Oliver Kennedy, Boris Glavic*, Sudeepa Roy+, Qitian Zeng*, Zhengjie Miao+
  * Illinois Institute of Technology
  + Duke University
- ❑ Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals https://link.springer.com/article/10.1023/A:1009726021843
- ❑ SeeDB:http://people.csail.mit.edu/mvartak/papers/seedb-full.pdf