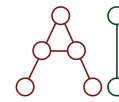


Uncovering Patterns in the LendingClub Dataset: An Exploration of Unsupervised and Supervised Techniques



Dataset Specifications

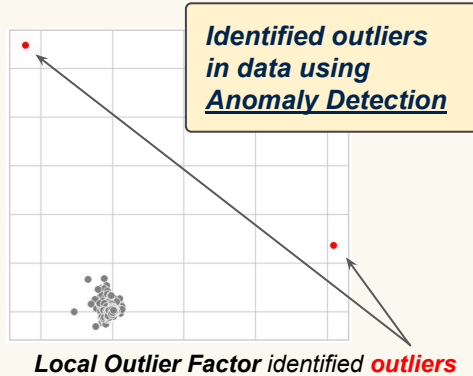
 **LendingClub**

Taken from [kaggle](#)

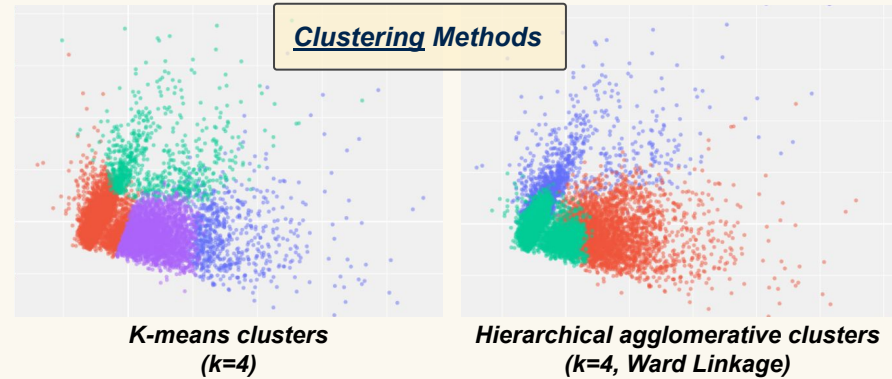
2.9M rows
2007-2020 Q3

142 cols

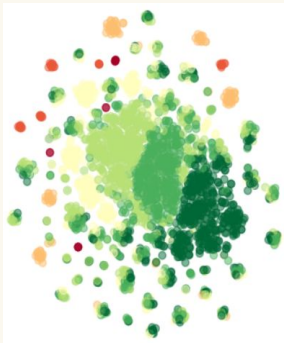
*Contains
Borrower
information,
loan details,
credit history*



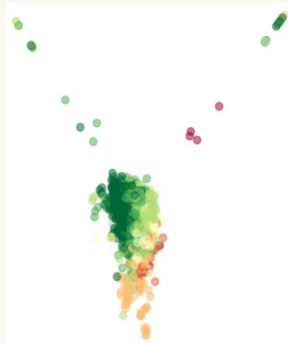
Clustering Methods



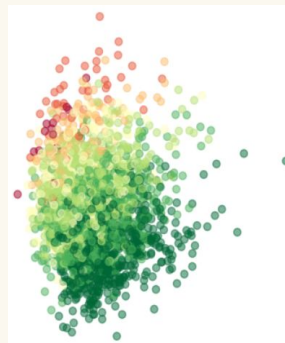
Dimensionality Reduction on LendingClub dataset features (color: risk grades)



t-SNE



Isomap



PCA

Binary Classification on delinquent loan profiles

Target Variable
Delinquent loan profile
(non-payment at least once in past 2 years)

Sample Features
Loan purpose
Employment length
Risk grade*
Home ownership type*
Loan term (36 or 60 months)*

*Top 3 in Feature Importance

9 LendingClub features used
AutoML with **PyCaret**
Cross-validated using 10-fold CV

	Model	Accuracy	AUC
xgboost	Extreme Gradient Boosting	0.8158	0.6238
rf	Random Forest Classifier	0.7791	0.5722
lr	Logistic Regression	0.5633	0.6127
ridge	Ridge Classifier	0.5583	0.0000
lda	Linear Discriminant Analysis	0.5583	0.6073

Best model: XGBoost