CIS432 - Predictive Analytics Using Python

Team Project #2

Decision Support System

MSBA2 - Team#16

Fangyuan Liu

Kaili Tan

Xuanhe Xu

Ze Long

Background Information

- **Goal:**
  - Evaluates the risk of Home Equity Line of Credit (HELOC)
- **Methodology:**
  - A predictive model and a decision support system (DSS)
- **Data info:**
  - 23 features used to predict whether an observation is paid as negotiated flag (good / bad)

Contents

01 Data Handling & Cleaning

1 Data Handling & Cleaning

Handling special values

- Drop all observations with -9
- Fill the -7 & -8 cells with median
  - Why median?
  - distribution, range, outliers
    -> mean is biased

Spliting into training & testing sets

- X_train, y_train (80% of all obs)
- X_test, y_test (20% of all obs)
  - For cross validation

Pre-processing categorical values

- MaxDelq2PublicRecLast12M
- MaxDelqEver
  - get dummies
  - combine the columns that representing same category

Scaling the value of features

- Standard scaling (for all features)
  - Why scaling?
  - to avoid bias on prediction

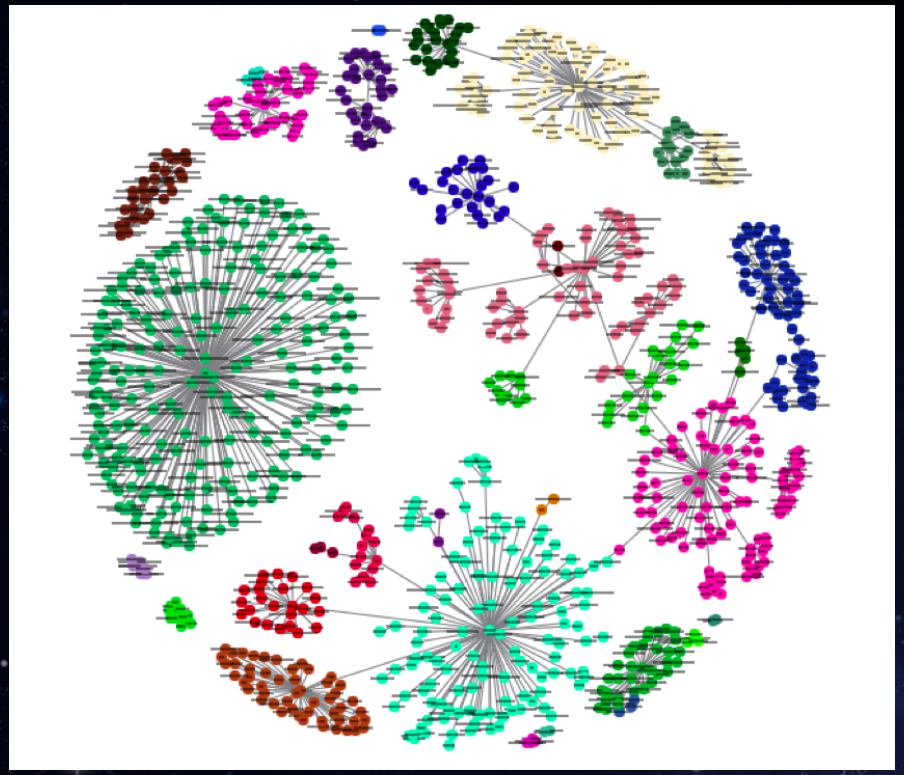
02

Modeling

- Models building
- Models evaluation
- Model selection

2 Modeling

Models Building - KNN



Search for parameters of the best KNN model using grid search, and get best parameters using training dataset:

- CV = 5, range: 1 – 23
- Best parameters:
  - n_neighbors: 22
  - Training accuracy = 0.7247

Apply to the testing dataset:

- KNN = 22
- Testing accuracy = 0.7192

2 Modeling

Models Building - SVM

Explore SVM using different kernels

- Linear
- Poly
- **rbf**

Search for parameters of the best SVM model with rbf kernel using grid search, and get best parameters using training dataset

- C: [0.1, 1, **10**, 100]
- gamma: [**0.001**, 0.01, 0.1]

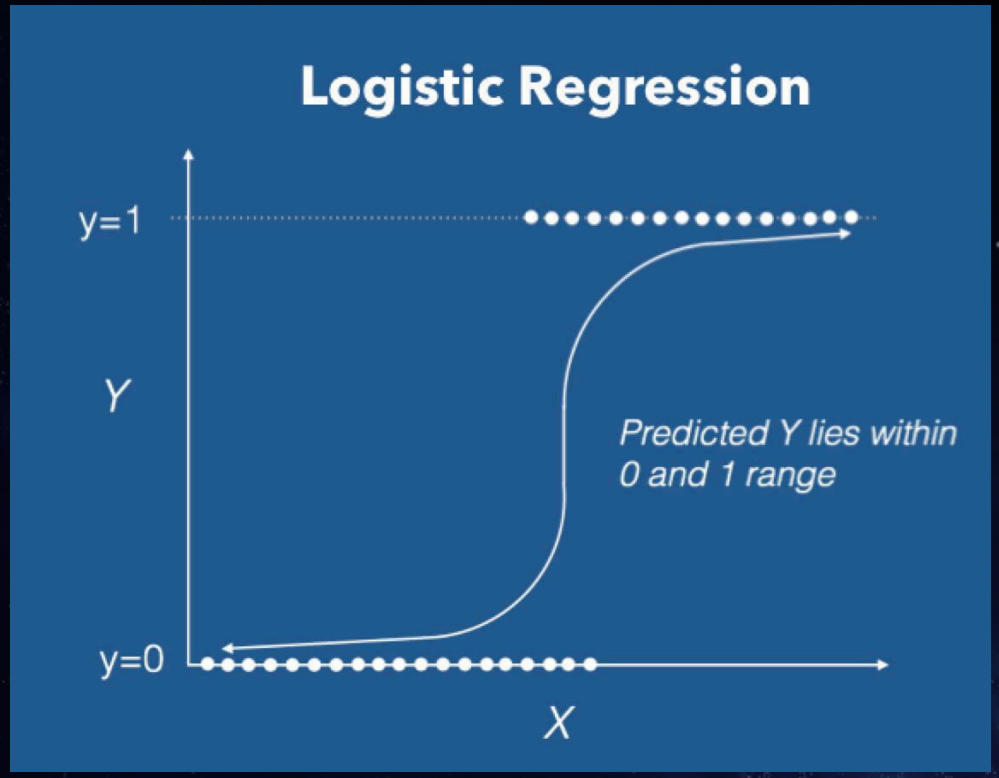
Apply the selected SVM model to testing dataset for validation

Testing accuracy

0.7344

2 Modeling

Models Building – Logistic Regression



Hyper-parameter

- C: np.logspace(-3, 3, 7)
- Penalty: ["l1", "l2"]

Best parameters

- C: 0.01
- Penalty: l2 (ridge)

Score

- 0.7182

1 Modeling

Models Building – Single Decision Tree

Hyper-parameter

Tree Depth: range(1, 6, 1)

Best parameter

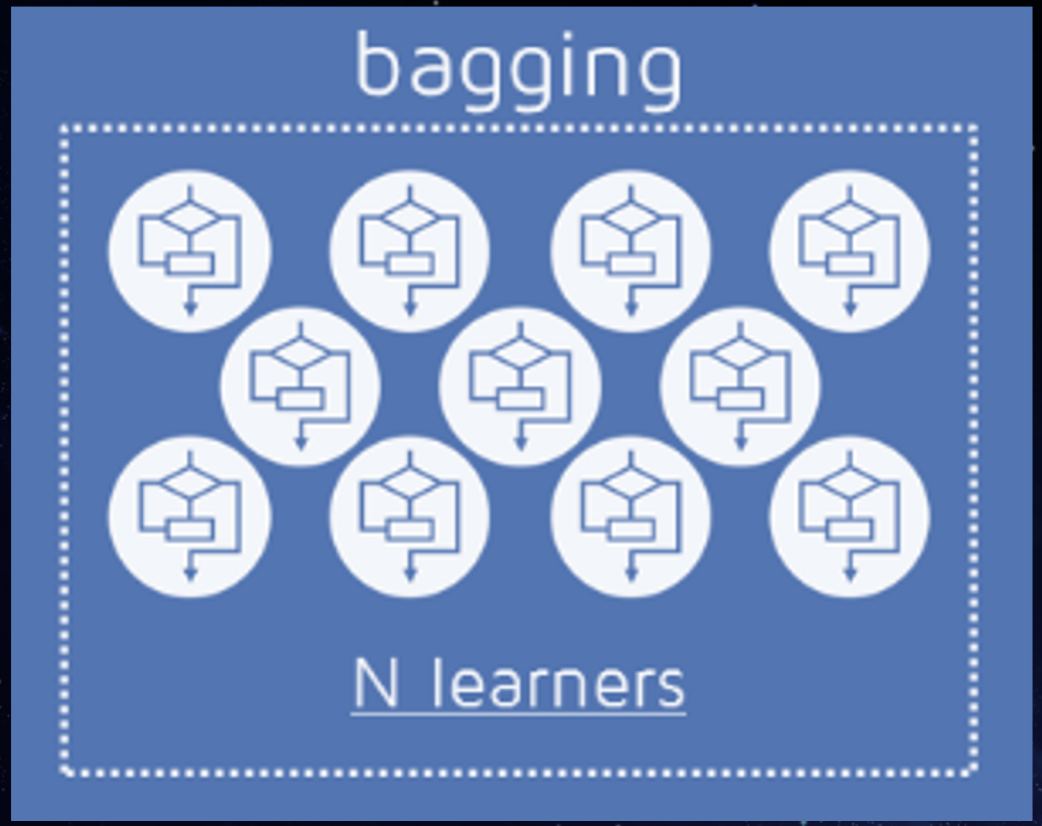
Tree Depth: 4

Score

0.7142

1 Modeling

Models Building – Bagging



Hyper-parameter

Tree Depth: range(1, 6, 1)

n_estimators: range(1, 40, 1)

Best parameter

Tree Depth: 5

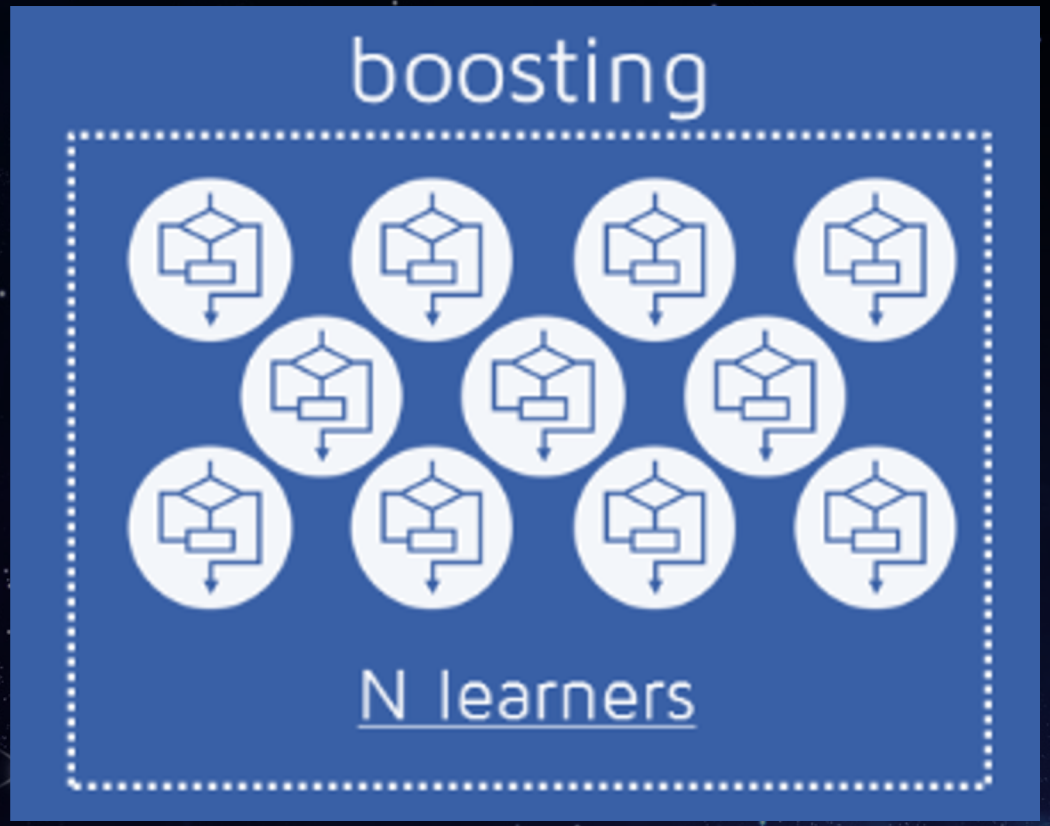
n_estimators: range: 33

Score

0.7272

1 Modeling

Models Building – Boosting



Hyper-parameter

n_estimators: [10, 20, 30, 40, 50]

Learning Rate: range(0,1, 1, 0.1)

Best parameter

Learning Rate: 0.9

n_estimators: range: 20

Score

0.7289

1 Modeling

Models Building – Random Forest

Hyper-parameter

Tree Depth: range(1, 6, 1)

n_estimators: range(1, 40, 1)

Best parameter

Tree Depth: 5

n_estimators: range: 38

Score

0.739782

2 Modeling

Models Evaluation & Selection

Main reference standard

- Accuracy (0 – 1)
- Simplicity & Interpretability

Results

- The best model: **random forest**
- The highest score: **0.74**

	model_names	model_score
0	Logistic regression	0.718196
1	KNN	0.719209
2	SVM	0.734415
3	Single decision tree	0.714227
4	Bagging	0.727207
5	Boosting	0.734915
6	Random forest	0.739783

03 Interactive Interface Demo

3 Interface

Clients Choices

- General glance of the model dataset
- Selection of individual records
- Choices of algorithm

Credit Risk Performance Model

Show dataframe

Choose a row of information in the dataset (0~10458):

5

which algorithm ?

Logistic regression

Logistic regression

KNN

SVM

Tree-based models

Boosting

RF

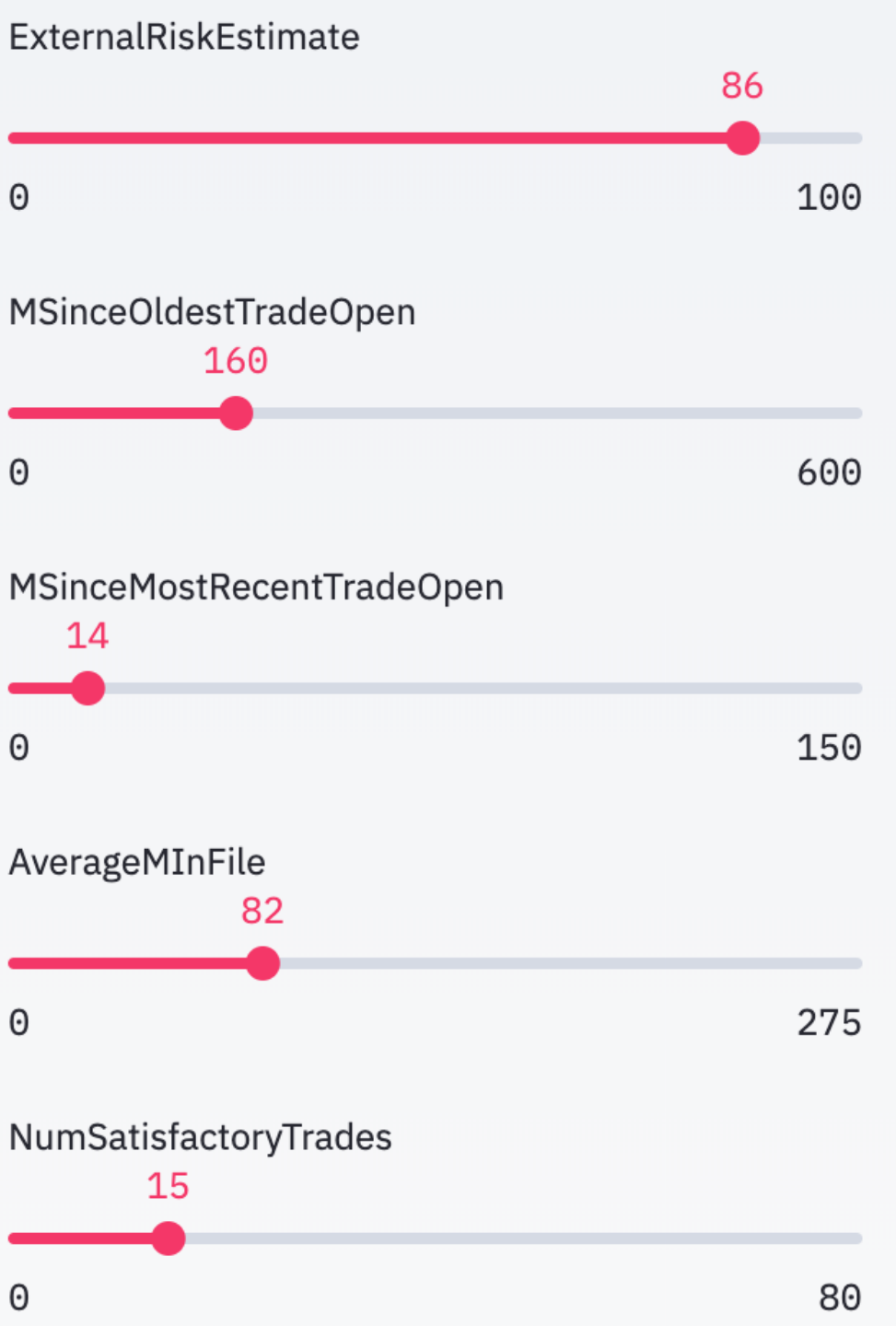
3 Interface

Inputs

Side bar:

variable of choice

Total 36 variables to modify personal inputs for prediction computation.



3 Interface

Outputs

- **prediction:**
  - The prediction outcome of credit performance based on the input data.
- **Accuracy:**
  - The accuracy of selected prediction method compare with the actual result.

Prediction: Good

Accuracy: 0.47237709072478457

Confusion Matrix:

	0	1
0	0	1041
1	0	932

04 Conclusion & Discussion

4 Conclusion & Discussion

Importance of data handling & cleaning

Importance of hyper-parameter selection

Importance of model evaluations & selections (considering the interpretability and overfitting)

Limitation 1:
Computing time

Limitation 2:
User friendly interface


Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehen-derit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.
EPS 10
ABSTRACT GRAPHIC
vector Illustration

THANK YOU