CIS 242

Final Project

Predictive Modelling and Decision Support System for Home Equity Line of Credit (HELOC)

Team Members:

Yi Lu,
Yibo Hu,
Yuanzhuo. Wang
Ziwei Shi

Apr. 2019

Agenda

I. Project Overview

1. Executive Summary
2. Data Insight

II. Data Cleaning

1. Special Value Cleaning
2. Categorial Variable Handling

III. Scaling & Model Selection

1. 5 Scaling Models & 7 Leaning Models
2. Model Comparison

IV. SVC: Decision Function

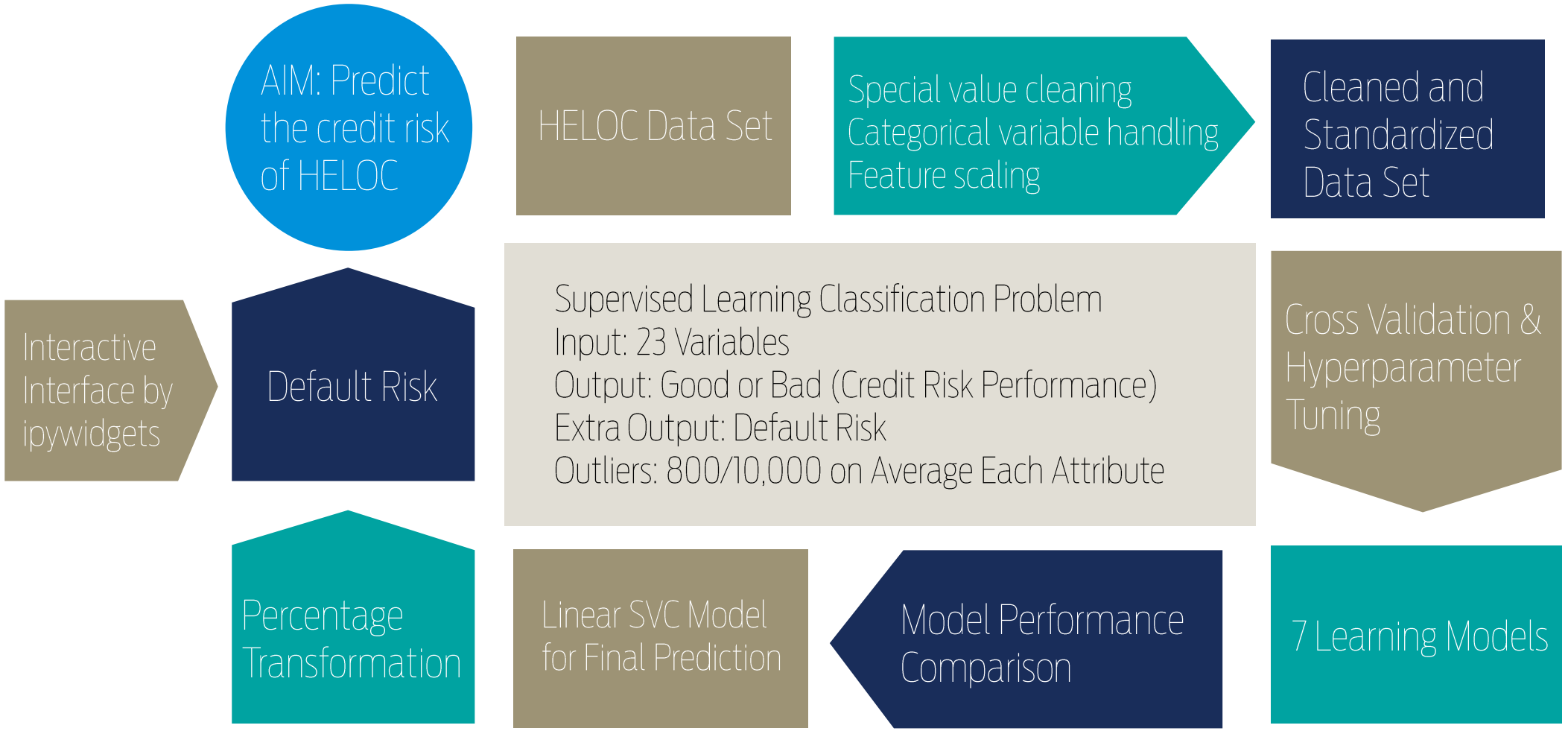
1. SVC Parameters
2. Linear Equation
3. Output

V. Interactive Interface & Explanation

1. Default Risk and Variable Weights
2. Demo

VI. Key Features

Project Overview & Data Insight



Data Cleaning

1. Drop Empty Rolls

- Drop 588 out of 10,459 rows with special value “-9”
- 323 of dropped applications were bad
- 256 of dropped applications were good

2. Frequency Count Table for Occurrence of Special Values

- Three special values “-7”, “-8”, “-9”
- -7 : condition not met (no info of the type)
- -8 : no useable/valid trades or inquiries (no usable info)
- -9 : no bureau record or no investigation *

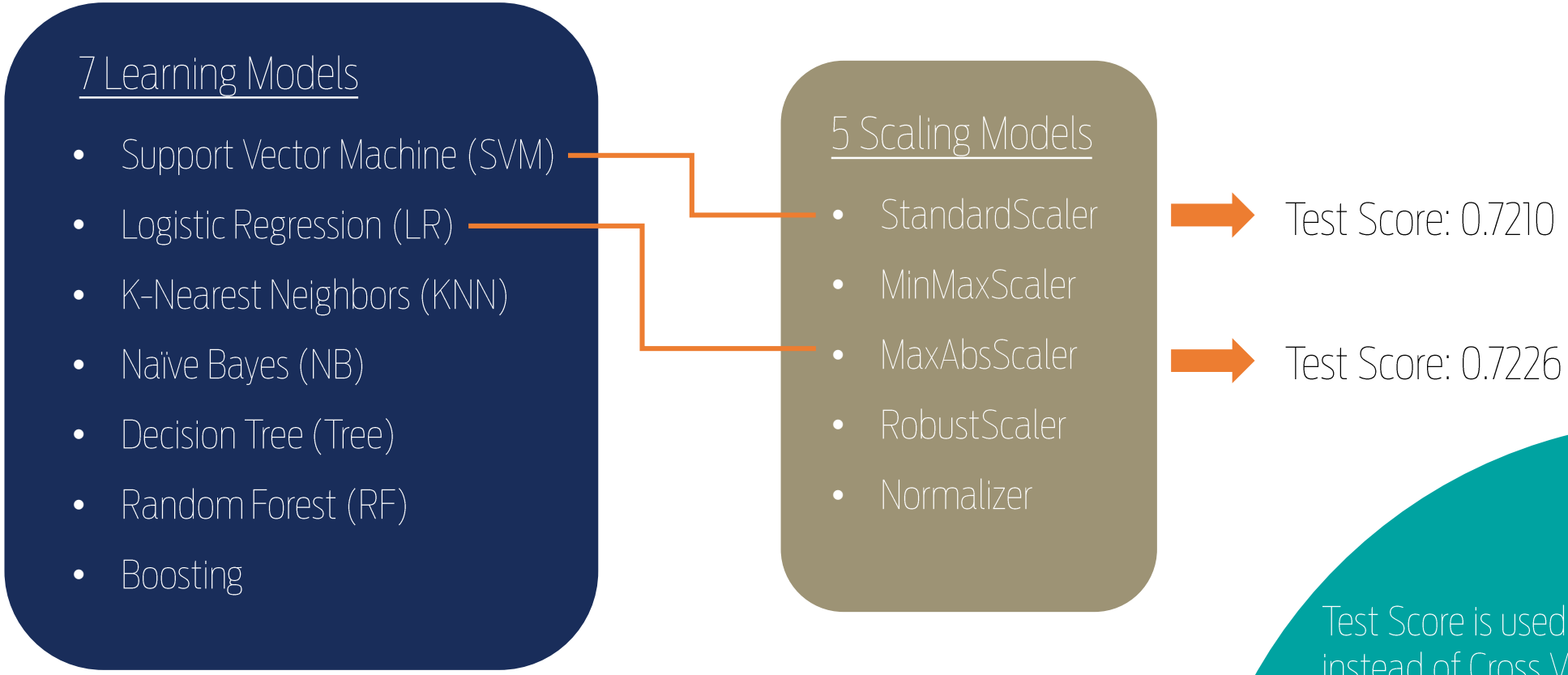
3. Clean the Special Value in Data Set

- -9: only 10 rows left; dropped rows with -9
- Mean and Median Methods
- Four combinations (median -7, median -8), (median -7, mean-8), (mean -7, median -8), (mean -7, mean -7)
- Median for -7 ; Mean for -8 (Highest Accuracy)

4. Convert and Store All Attributes to Indicator Variables

- Convert all Categorical Variables to Dummy Variables
- Converted Dummy Variables has binary values

Scaling and Model Selection



Test Score is used here instead of Cross Validation (CV) due to the large dataset we have. Test Score can be more accurate in this scenario

Scaling and Model Selection

Two Combinations with Highest Test Score After Hyperparameter Tuning:

- SVM with StandardScaler
  Test Score: 0.7210
- LR with MaxAbsScaler
  Test Score: 0.7226

SVM VS. LR

- SVM minimizes hinge loss
- LR minimizes logistic loss
- Logistic loss diverges faster than hinge loss
- LR will be more sensitive to outliers
- In our scenario, we have a large number of outliers (800 per attribute)
- Therefore, we use SVM as the classifier for final decision model

SVC: Decision Function

Final SVC Parameters:

- [0.1, 1] used as penalty
- "linear" as kernel

Linear Equation and Calculating Functions:

- $Y = a + b\boldsymbol{x_1} + c\boldsymbol{x_2} + d\boldsymbol{x_3} + \ldots$
- *X* in this equation is standardized input variable
- "final_best_model.coef()" finds *b, c, d, ….*
- "final_best_model.intercept()" finds *a*
- "final_model.decision_function()" finds *Y*

Output and Explanation:

- Y is the final output indicates the risk of loans
- Y (variable "result") mostly ranges in (-5, 5)
- Positive Y means Good risk performance
- Negative Y means Bad risk performance
- More positive Y means Better risk performance
- More negative Y means Worse risk performance

→ Can the output be quantified?

Interactive Interface & Explanation

User Interface Model:

- Ipywidgets
- No coding background required for users
- Input variables by dragging the bar
- Output in Default Risk
- 5 major contributing attributes to the output will be shown

Default Risk:

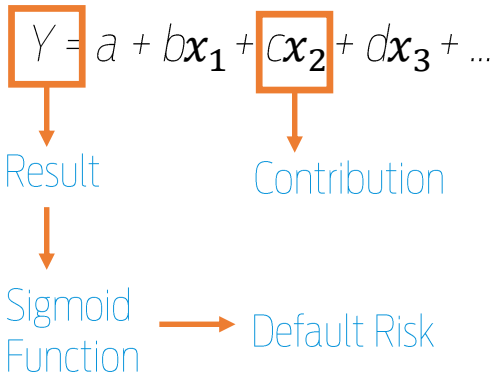
- The chance that companies or individuals will not make the required payment
- Calculated by inputting Y into sigmoid function
- Range from 0% to 100%

Contribution:

- Each attribute has a contribution rating to the output
- 5 attributes will be in the order of the extent of affection
- The contribution is calculated by

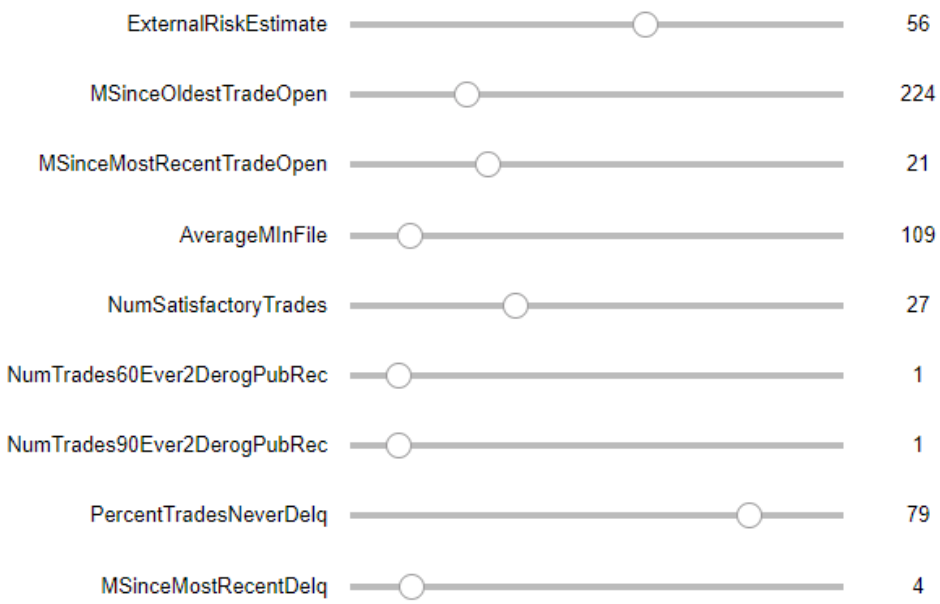
  Standardized input value * model coefficient

Equation Explanation:



Demos

Part of Input Interface:



Final Output Example:

```
# Demo 1: Good
input_demo(10)
test_demo(5)
```

```
Decision function result: 0.5539094320020761
Default Risk: 36.50%
Good

Key attributes that determine the prediction result:
[Order by importance]
```

	Order of Importance	Attribute	Value	Contribution in Decision
0	1.	ExternalRiskEstimate	82	0.408455
1	2.	NumInqLast6M	0	0.405448
2	3.	NetFractionRevolvingBurden	15	0.180792
3	4.	MaxDelq2PublicRecLast12M	7	0.124925
4	5.	NumBank2NatlTradesWHighUtilization	0	0.106290

Key Features

- Using tables to gain data insights into special values.
- Data cleaning methods of handling special values.
- Hyper-parameter tuning to find the best model by comparing test scores.
- Difference between Linear SVC and LR—sensitivity to outliers.
- Interactive interface by “Ipywidgets”.
- Explanations of our result by giving users a probability of Default Risk and 5 most significant contributing factors based on different user inputs.

Thank You !

Apr. 2019

Here is so much blank
IDK what to write here
Yaron is a good professor
You guys are good
Good luck with finals

First Class
Yaron: “It’s ok if you have no coding background. I’ll go over the basics”

Then finished CIS 191 in 2 and half hours

PowerPoint Design: Ziwei Shi