Kaggle:
Santander
Customer Transaction Prediction

Student: Yunjie Cao

Instructor: Pierre Nugues

Date: 05/29/2019

Agenda

- Background
- Data
- Model
- Result
- Discussion

Background

- Finance and business development
- Computation capability
- Machine learning methods
- Automatic algorithms in business scenarios

Santander customer transaction prediction

- Is a customer satisfied?
- Will a customer buy this product?
- Can a customer pay this loan?

-> Will they make the transaction?

Binary classification problem

Data

200 features represent every customer

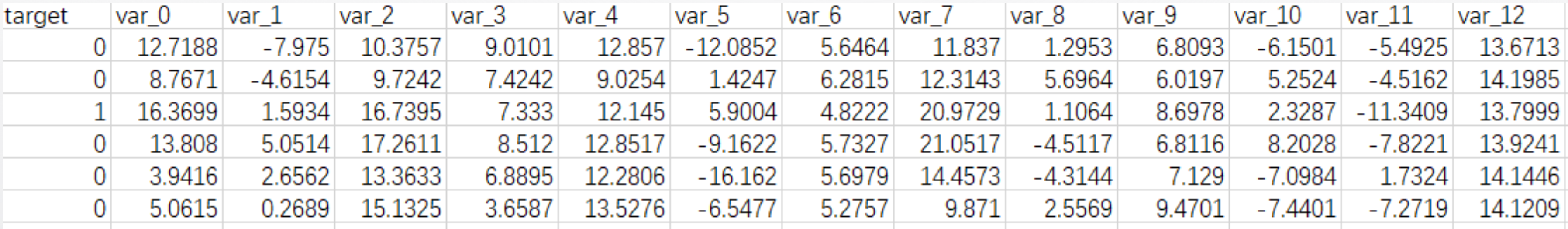
TrainSet:

200000 customers 200 features + 1 label

TestSet:

200000 customers 200 features -> predict

Data



target	var_0	var_1	var_2	var_3	var_4	var_5	var_6	var_7	var_8	var_9	var_10	var_11	var_12
0	12.7188	-7.975	10.3757	9.0101	12.857	-12.0852	5.6464	11.837	1.2953	6.8093	-6.1501	-5.4925	13.6713
0	8.7671	-4.6154	9.7242	7.4242	9.0254	1.4247	6.2815	12.3143	5.6964	6.0197	5.2524	-4.5162	14.1985
1	16.3699	1.5934	16.7395	7.333	12.145	5.9004	4.8222	20.9729	1.1064	8.6978	2.3287	-11.3409	13.7999
0	13.808	5.0514	17.2611	8.512	12.8517	-9.1622	5.7327	21.0517	-4.5117	6.8116	8.2028	-7.8221	13.9241
0	3.9416	2.6562	13.3633	6.8895	12.2806	-16.162	5.6979	14.4573	-4.3144	7.129	-7.0984	1.7324	14.1446
0	5.0615	0.2689	15.1325	3.6587	13.5276	-6.5477	5.2757	9.871	2.5569	9.4701	-7.4401	-7.2719	14.1209

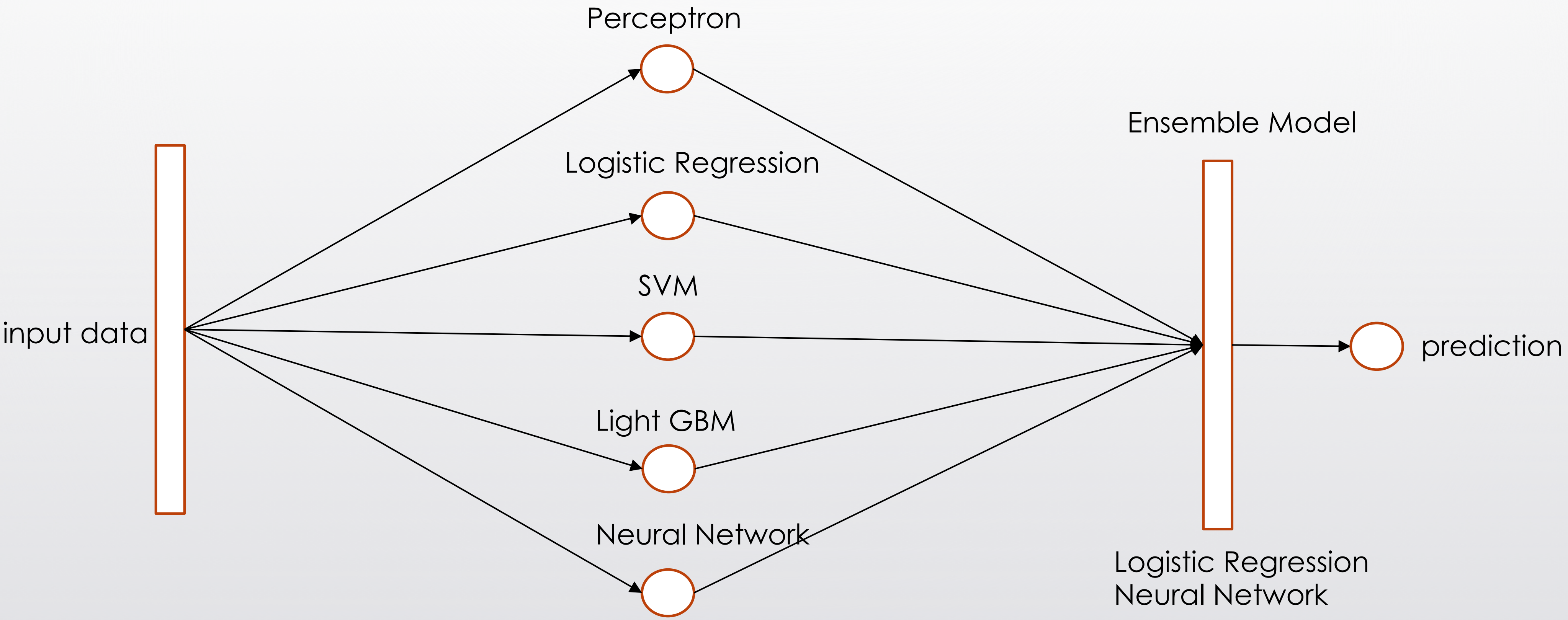
Unbalanced data:

negative:positve 9:1 -> sample

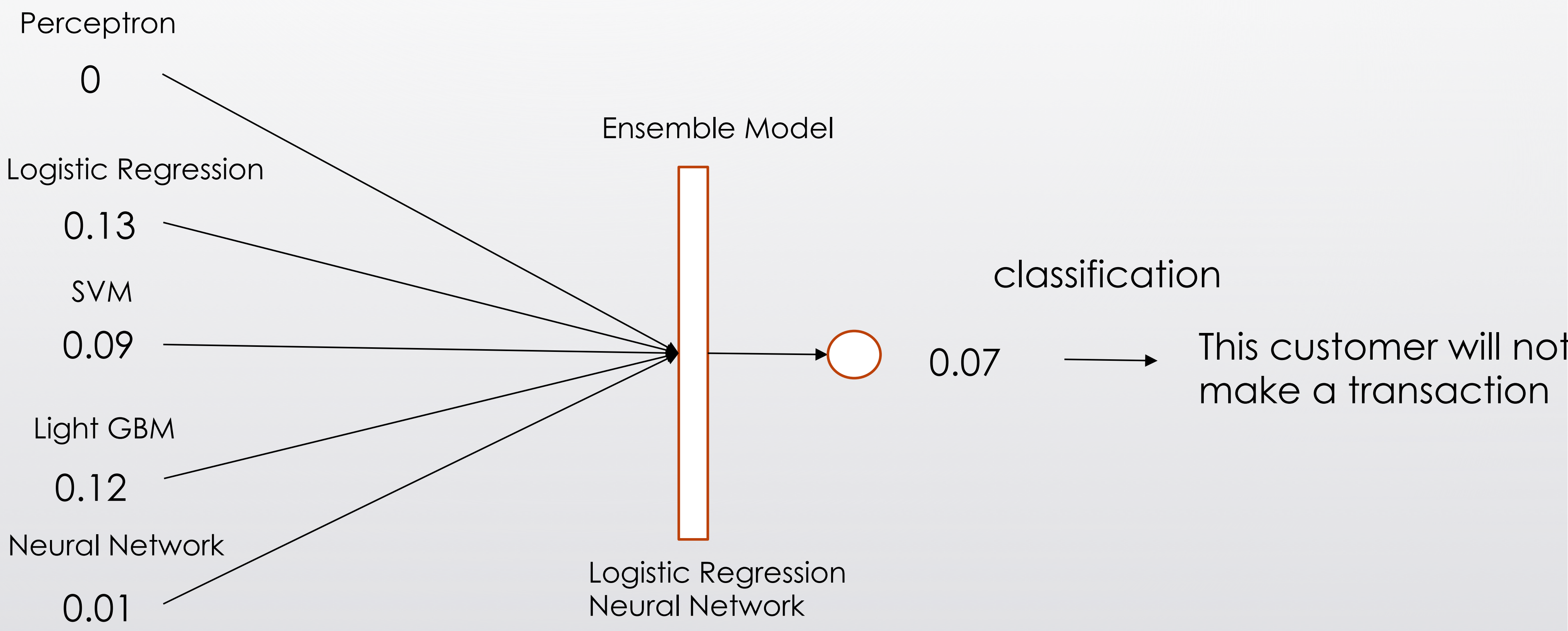
Feature engineering :sum, min, max, mean, std, skew, kurt, med

Normalization

Model


Perceptron
Logistic Regression
SVM
Light GBM
Neural Network
input data
Ensemble Model
Logistic Regression
Neural Network
prediction

Model



Perceptron

0

Logistic Regression

0.13

SVM

0.09

Light GBM

0.12

Neural Network

0.01

Ensemble Model

Logistic Regression
Neural Network

classification

0.07

This customer will not make a transaction

Result

- Single Model:

Algorithm	Accuracy
Perceptron	0.732
Logistic Regression	0.773
Support Vector Machine	0.861
Light GBM	0.880
Neural Network	0.891

- Ensemble Model (Light GBM + SVM + Neural Network)

Ensemble Algorithm	Accuracy
Logistic Regression	0.897
Neural Network	0.896

Discussion

- My Best Accuracy: 0.897
- Best Accuracy: 0.925

Problems:

Fake samples in train data

Thank you!