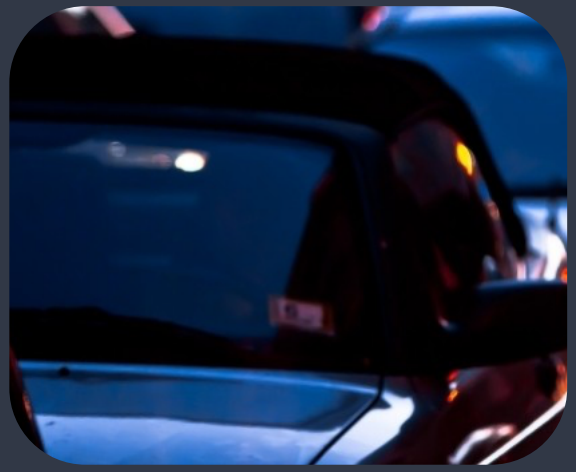


Car Insurance Plan

Machine Learning I – Final Project

Lianjie Shan, Xi Zhang

IDX

CONTENTS

Background

HOW DOES MACHINE LEARNING WORK ON CAR INSURANCE?

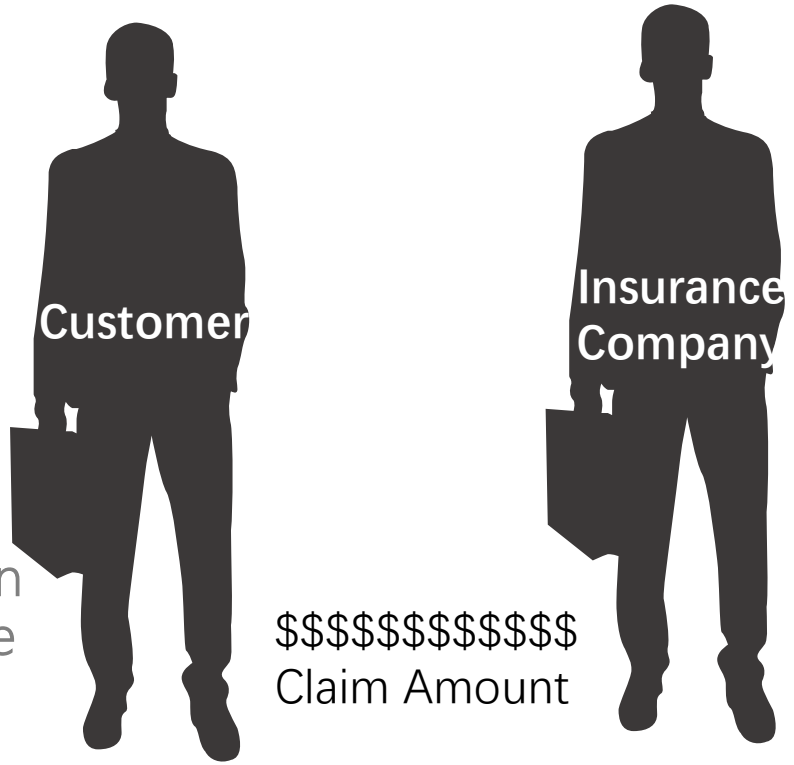
Car Insurance Market

Self Assessment

Possibility of Customer Claim

Insurance Plan Selection

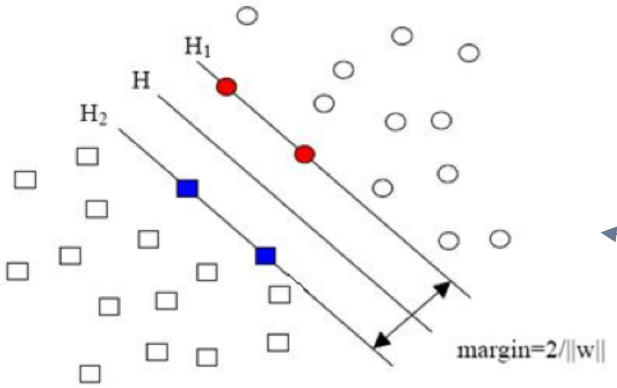
Prediction of Claim Amount



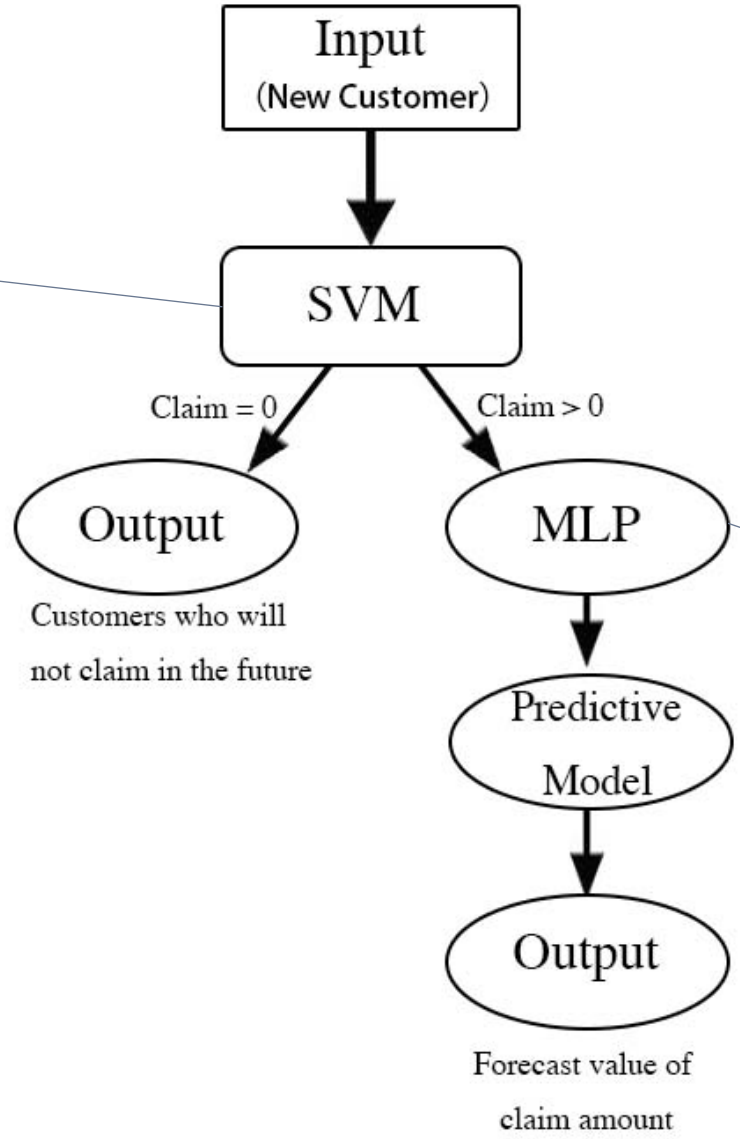
Most Suitable Plan with Lowest Price

$$$$$$$$$$$
Claim Amount

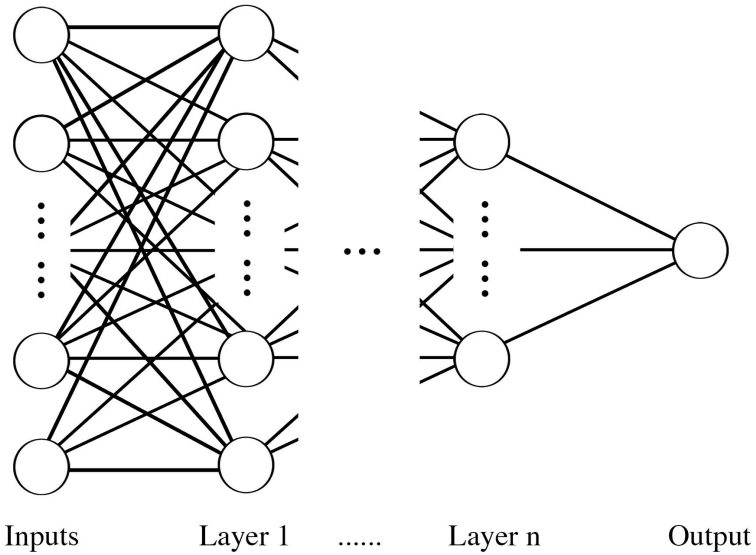
The Highest Profit

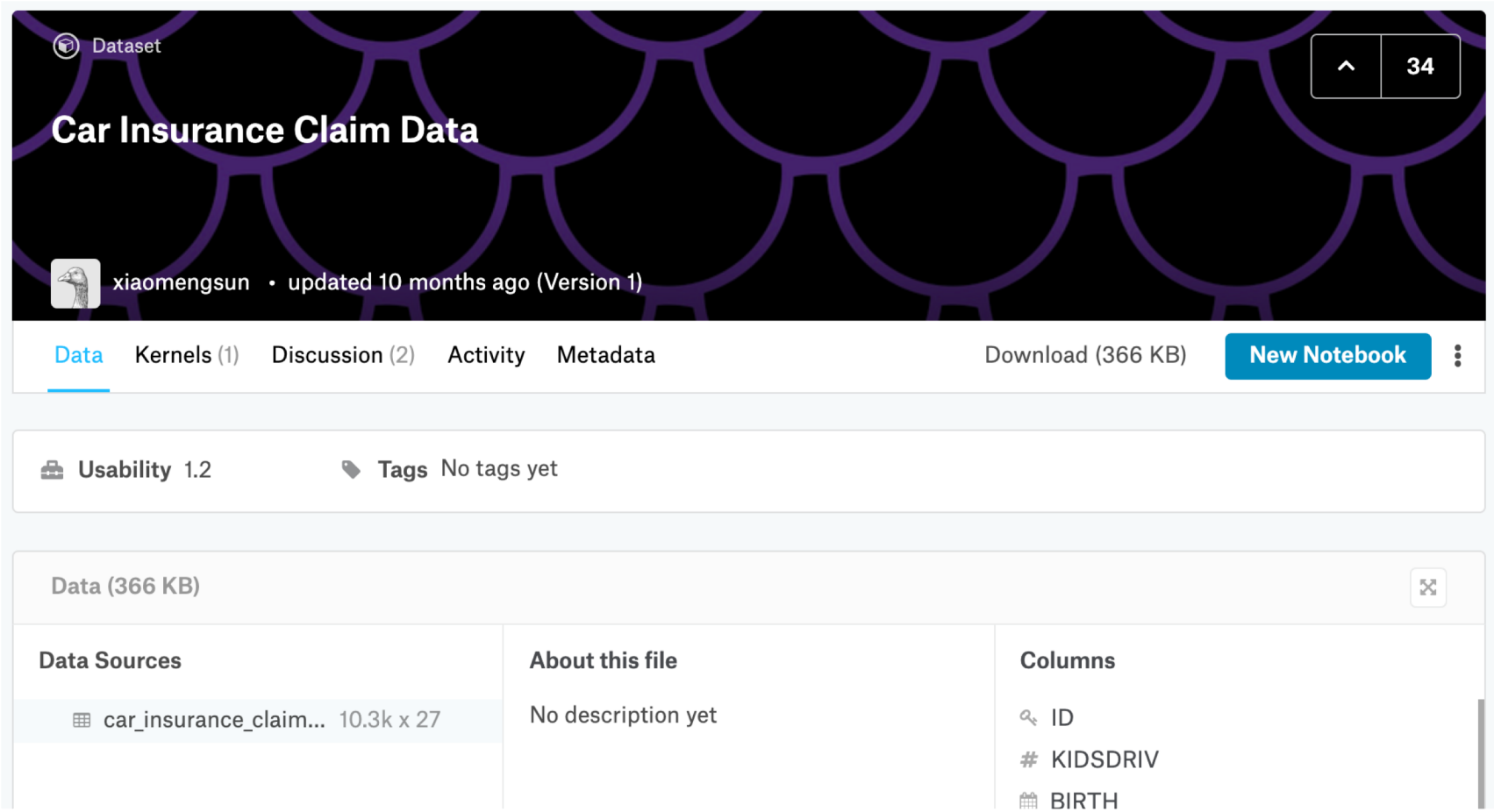


SVM
classify claim (1)
and
non-claim customers (0)



MLP
Predict specific
amount of claims



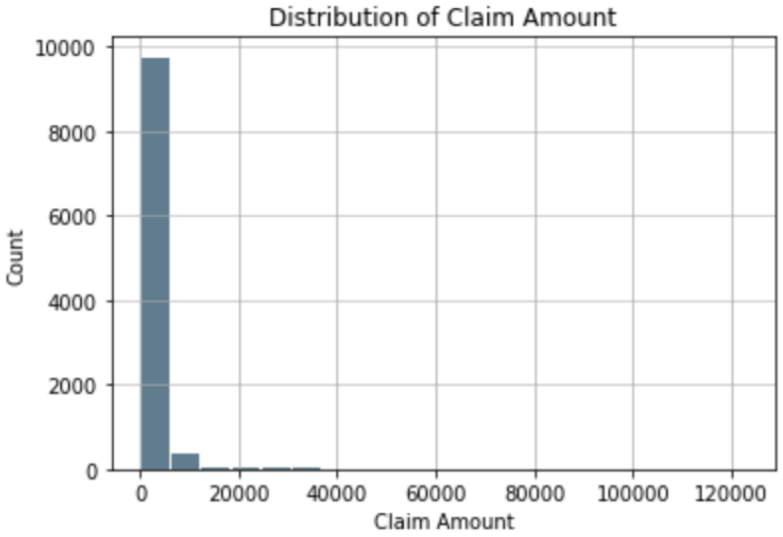


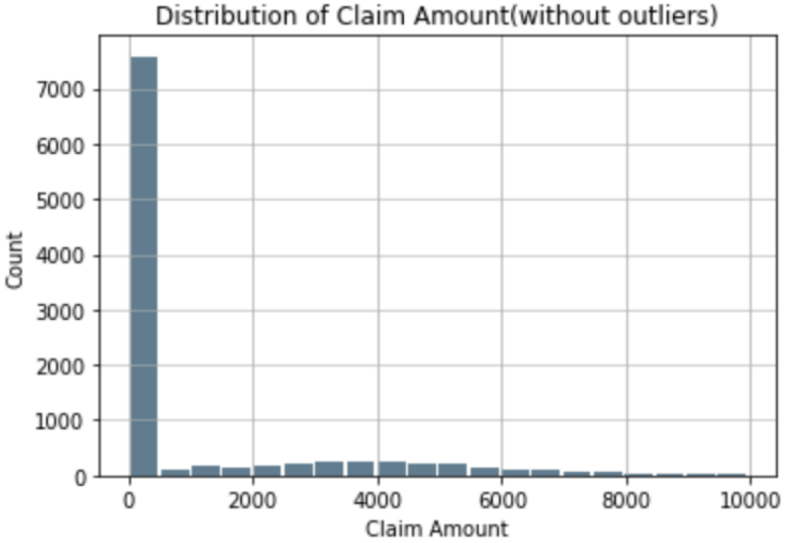
Data Resource

Observations: 10,302

Variables: 27

ID
KIDSDRIV
BIRTH
AGE
HOMEKIDS
YOJ
INCOME
PARENT1
HOME_VAL
MSTATUS
GENDER
EDUCATION
OCCUPATION
TRAVTIME
CAR_USE
BLUEBOOK
TIF
CAR_TYPE
RED_CAR
OLDCLAIM
CLM_FREQ
REVOKED
MVR_PTS
CLM_AMT
CAR_AGE
CLAIM_FLAG
URBANICITY





01 Remove Useless Columns

02 Digitalize Categorical Data

03 Unify Data Types

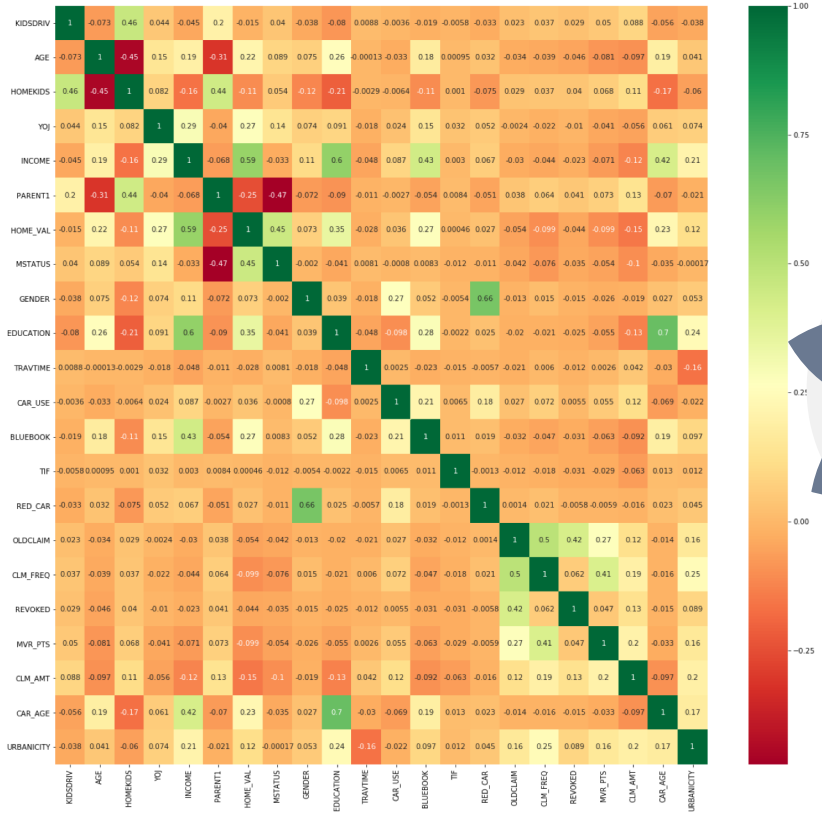
04 Remove Outliers

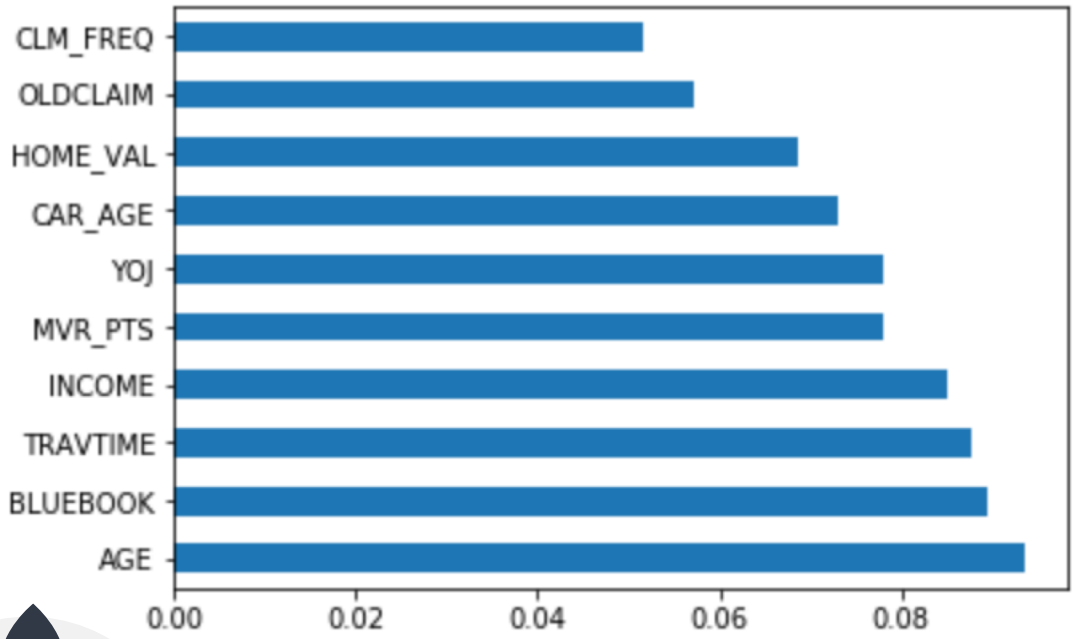
05 Remove Missing Values

Observations: 10,302 → 8,014

Correlation Plot

Relation Overview





Decision Tree

Top 10 Importance Features

Select top5 features:–

‘AGE’ ,’ BLUEBOOK’ , ‘TRAVTIME’ , ‘INCOME’ , ‘MVR_PTS’

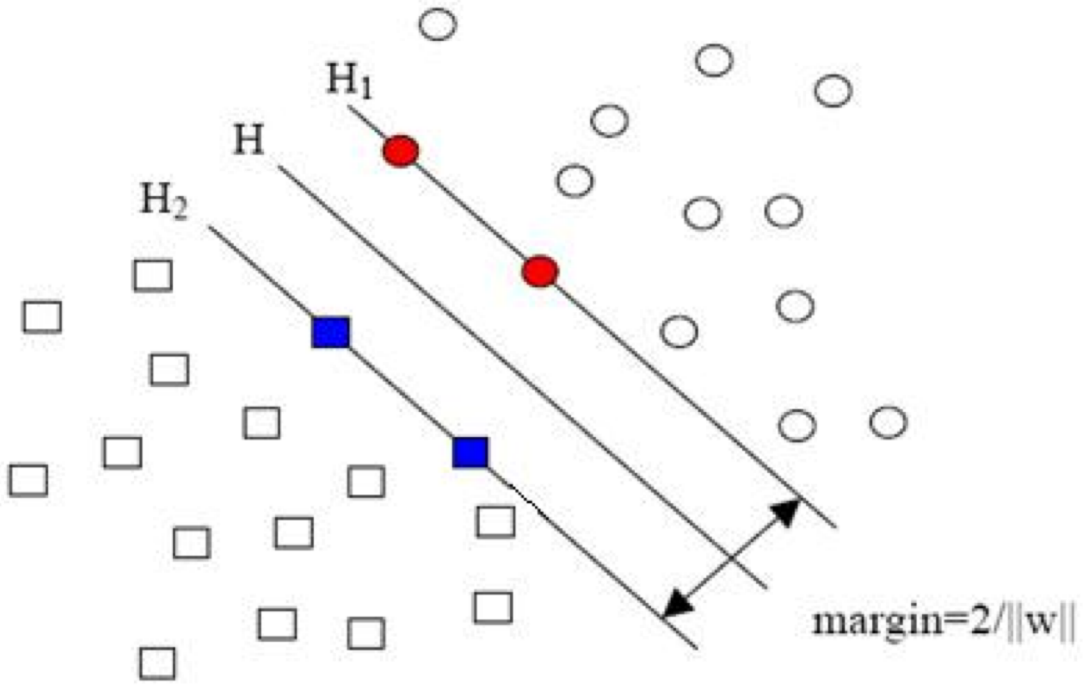
SVM

DIVIDE CUSTOMERS BY WHETHER THEY WILL CLAIM

Principle

Support Vector Method (SVM) as a popular machine learning tool is most used for Classification.

SVM tries to find a plane that has the maximum margin and the maximum distance between data points of both classes.



Data preparation

	BLUEBOOK	TRAVTIME	INCOME	MVR_PTS	AGE	CLM_AMT
0	14230.0	14.0	67349.0	3.0	60.0	0.0
1	14940.0	22.0	91449.0	0.0	43.0	0.0
2	21970.0	26.0	52881.0	2.0	48.0	0.0
3	4010.0	5.0	16039.0	3.0	35.0	0.0
5	18000.0	36.0	114986.0	3.0	50.0	0.0
6	17430.0	46.0	125301.0	0.0	34.0	2946.0

	BLUEBOOK	TRAVTIME	INCOME	MVR_PTS	AGE	CLM_AMT
0	14230.0	14.0	67349.0	3.0	60.0	0.0
1	14940.0	22.0	91449.0	0.0	43.0	0.0
2	21970.0	26.0	52881.0	2.0	48.0	0.0
3	4010.0	5.0	16039.0	3.0	35.0	0.0
5	18000.0	36.0	114986.0	3.0	50.0	0.0
6	17430.0	46.0	125301.0	0.0	34.0	1.0

Dataset transformation

Replace all claim values greater than 1 with 1

Dataset segmentation

divided data into training data (70%) and test data (30%).

Model Implement

Kernel selection : linear kernel

$$k(x, y) = x^T y + c$$

The principle of linear kernel is that the output is given by the inner product plus an optional constant c.

Results and Evaluation

After SVM model was built, training dataset was inputted into Python to get the final classification result

```
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.93      0.84      2135
           1       0.44      0.15      0.22       727

   micro avg       0.74      0.74      0.74      2862
   macro avg       0.60      0.54      0.53      2862
weighted avg       0.68      0.74      0.68      2862


Accuracy :  73.51502445842068
```

MLP

PREDICT SPECIFIC CLAIM AMOUNT

Data preparation

Dataset selection

The value of claim is bigger than one

standardization

eliminate dimensional influence between features

Dataset segmentation

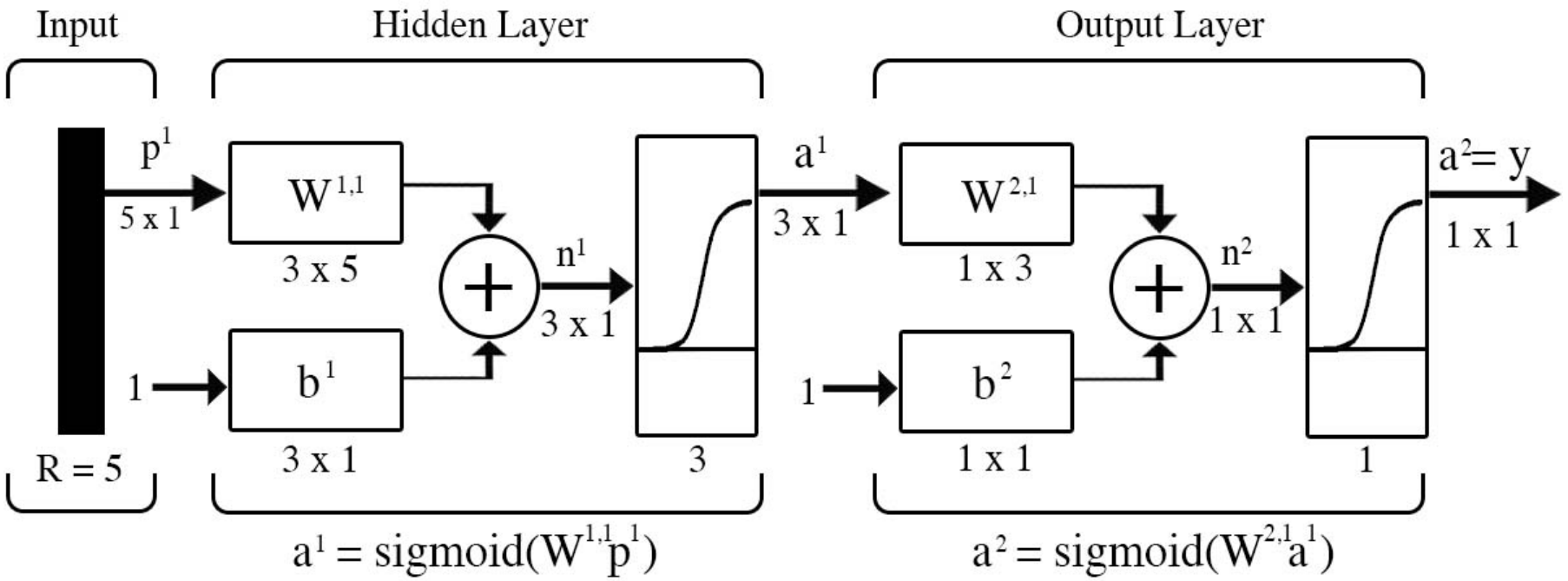
divided data into training data (70%) and test data (30%).

Data preparation

	BLUEBOOK	TRAVTIME	INCOME	MVR_PTS	AGE	CLM_AMT
6	0.074086	0.467391	0.000000	0.230769	0.483333	0.291478
8	0.118703	0.260870	0.137686	0.000000	0.200000	0.688553
10	0.161179	0.119565	0.118515	0.000000	0.733333	0.526172
11	0.140435	0.228261	0.097696	0.000000	0.266667	0.449420
13	0.176655	0.163043	0.053070	1.000000	0.400000	0.612204

After standardization, 403 values of training data can be obtained, and 174 values of test data can be obtained.

Network Architecture



From the architecture, the number of layer is 2; the dimension of the feature is 5; the number of neuron is 3; the transform function is log sigmoid.

Building Predictive Model

```
Stage 1) Random starting synaptic weights:
    Layer 1 (3 neurons, each with 5 inputs):
[[-0.16595599  0.44064899 -0.99977125]
 [-0.39533485 -0.70648822 -0.81532281]
 [-0.62747958 -0.30887855 -0.20646505]
 [ 0.07763347 -0.16161097  0.370439  ]
 [-0.5910955   0.75623487 -0.94522481]]
    Layer 2 (1 neuron, with 3 inputs):
[[ 0.34093502]
 [-0.1653904 ]
 [ 0.11737966]]
Stage 2) New synaptic weights after training:
    Layer 1 (3 neurons, each with 5 inputs):
[[-3.23330008  0.5400038  -2.98145774]
 [ 0.02438085 -5.66844489 -1.64677393]
 [ 1.66498726 -4.3400436   2.81042168]
 [-4.39565522  5.48878752 -9.22471803]
 [-0.72170302 -5.26456209 -7.29046498]]
    Layer 2 (1 neuron, with 3 inputs):
[[-2.52121198]
 [-1.30651473]
 [ 6.79308023]]
```

After training the dataset, weight vector was produced and the preliminary model was built.

Dataset inversion

[0.3673214]		[3671.99163825]
[0.42892033]		[4282.74509204]
[0.38475559]		[3844.85163628]
[0.37078258]		[3706.30930536]
[0.32756746]		[3277.8313754]
[0.38619124]		[3859.08617053]
[0.37190939]		[3717.48156009]
[0.44247311]		[4417.12087744]
[0.39065171]		[3903.31168905]
[0.37662705]	→	[3764.25721683]
[0.37076745]		[3706.15926393]
[0.37946452]		[3792.39073447]
[0.3826891]		[3824.36247572]
[0.36139664]		[3613.24770296]
[0.39366315]		[3933.17009324]
[0.37221891]		[3720.55045839]
[0.32236985]		[3226.29706708]
[0.31275087]		[3130.92492173]
[0.43665823]		[4359.46630453]
.		.
.		.
.		.

With the purpose of intuitively observing the predictive results. Using the inverse function in Python to inverse prediction values from standardized to real.

Evaluation of MLP Model

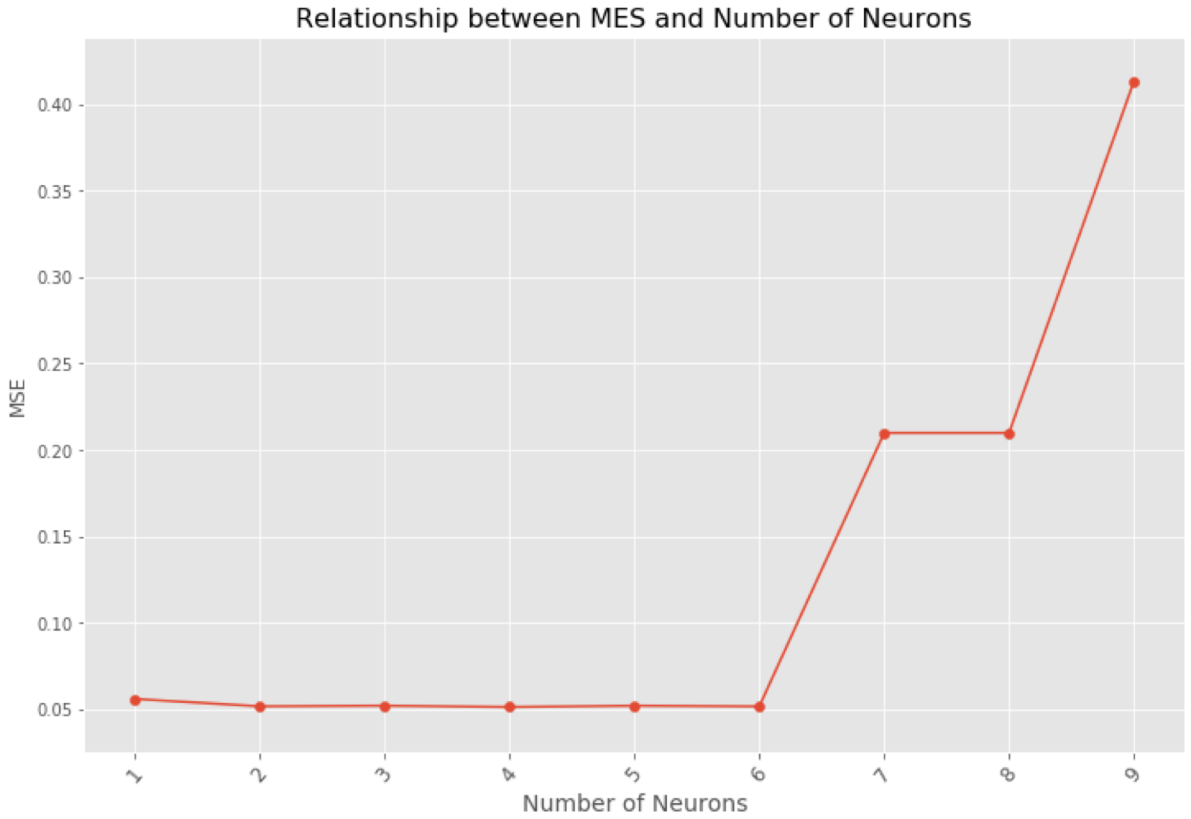
The mean square error(MSE), as a popular method of calculating errors, is often used as a standard to evaluate a model.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2$$

In order to evaluate the predictive model, we use the test dataset to calculate the value of MSE and the value of MSE is equal to 0.3

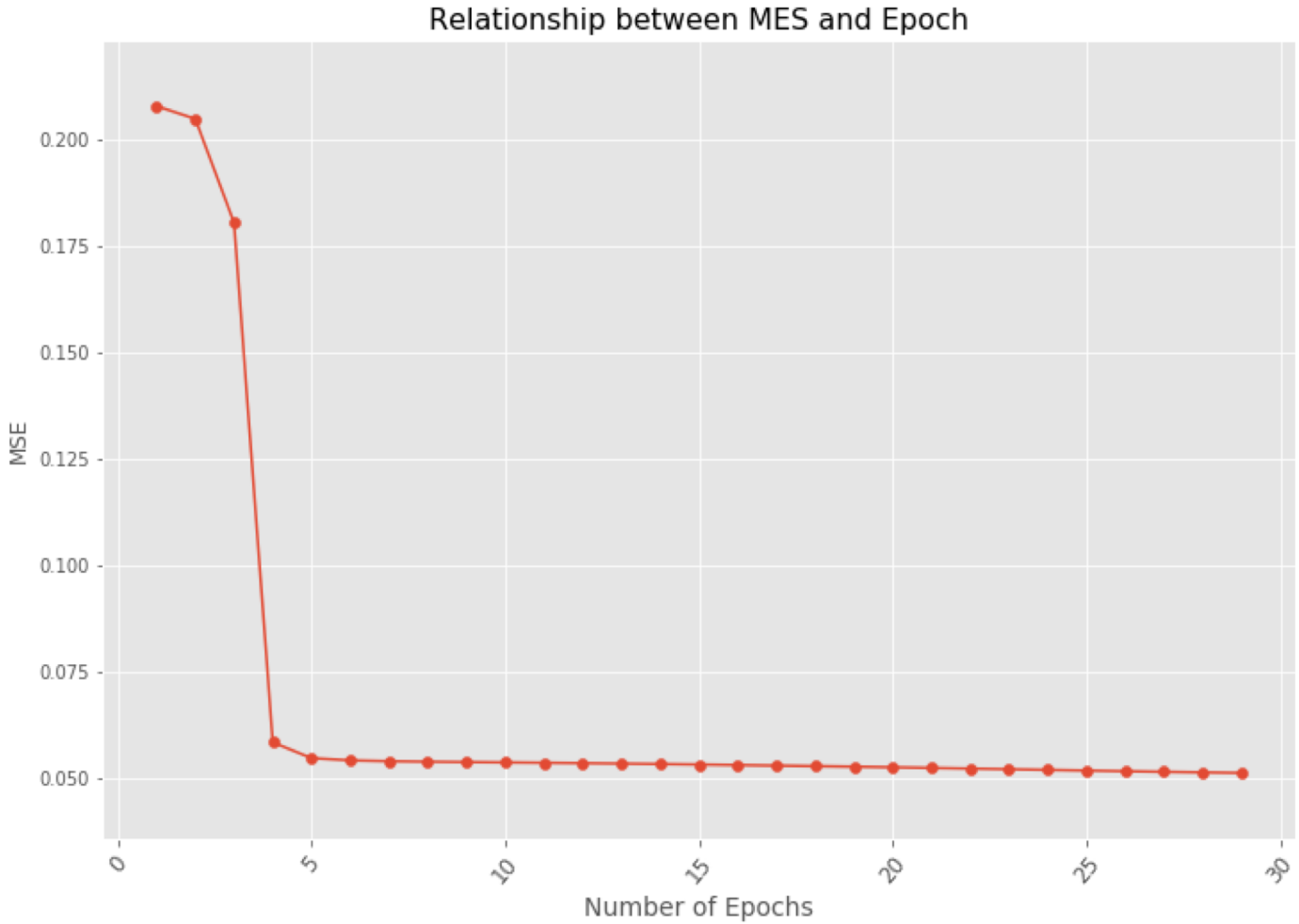
Improvement of MLP Model

After building the predictive model, adjusting the number of neurons and epochs to improve the accuracy of the MLP model.



From this plot, when the value of neurons is greater than 6, an obvious increase in the value of MSE. And the ideal number of neurons is 3.

Improvement of MLP Model



From this plot, as the epoch was greater than 5, the MSE value became minimal and stable. Thus, the optimal value of epoch is equal to 5.

Comparison of MSE Values

	The initialization of the neurons and epochs (neurons = 5; epochs = 1000)	The optimal number of neurons and epochs (neurons = 3; epochs = 5)
The value of MSE	0.3	0.03

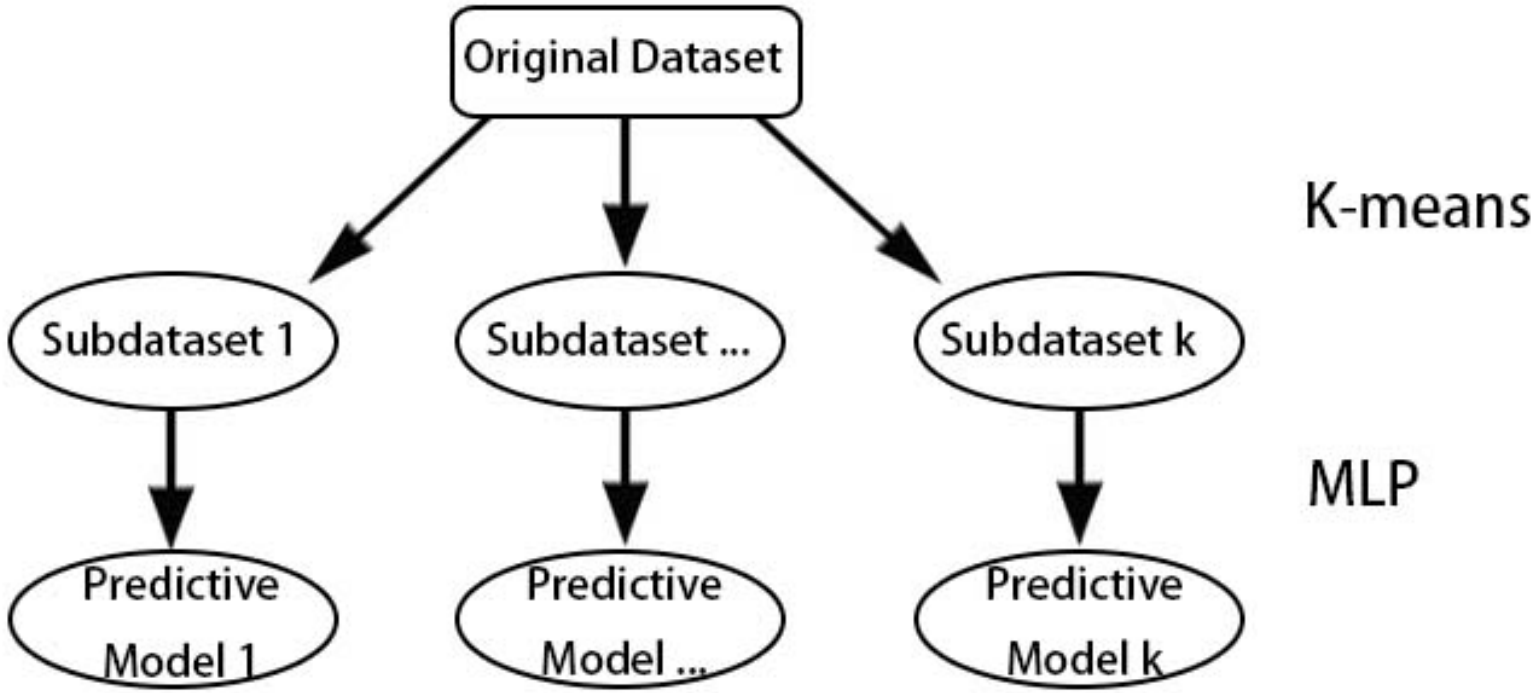
From the table above, the value of MSE is about 0.03 which is a significant decline compared with before. Thus, the accuracy of the model has been Improved.

Improvement

APPLICATION OF K-MEANS ALGORITHM

K-means algorithm

Due to the huge and complex data volume of users in real life, the accuracy of our prediction model will be reduced to some extent. In this case, we use K-means algorithm to classify our original data and use the MLP method for each subclass to build predictive model.



Thanks For Watching