FIT 5142 advanced data mining

Predict Geographic Average Income by Using Supervised Learn Algorithm

Name: Fanchao Kong

Student Number: 27946789

List

1 Introduction

These day, data mining play an important role in many field such as industry, research and medicine etc. It help people discover knowledge and value from data. The data set *Insightful & Vast USA Statistics*[3], contain 80 attributes related to Geographical location's finance condition, education condition and citizen detail such age and marriage and it has 39000 instances. This report are focus on the relationship among these attribute. By mining this data set step by step, the relationships among these attributes could be clear. The relationships contain huge value for both business and society. For example, it is able to predict family income condition by using these attribute.

The structure of this report will be split to six parts, which are introduction, literature review, data exploration, experimental design, result analysis and conclusion. The previous data mining case will be summarized, and the useful method from these case will be used for references at literature part. In the data exploration part, will focus data preprocessing about how to make data format and content better, in that the model for analyzing will be work more effectively and efficiently, and some wrangling details will be represented at here. For experimental design, this part will explain the algorithms for analyzing data. The algorithm could be classification or cluster method. After analyzing data, the result will be researched at here, the relationship between each attribute should be clear. As last, the conclusion about this report will be summarized.

2 literature review

This part we look the data mining methods that other people do before, and extract advantages, which could be using in this data set researching. Literature mainly has three parts.

- Previous work in this data set
  This part focus on how does other people handle this data set and get insight from them. The content is like combination of others work.

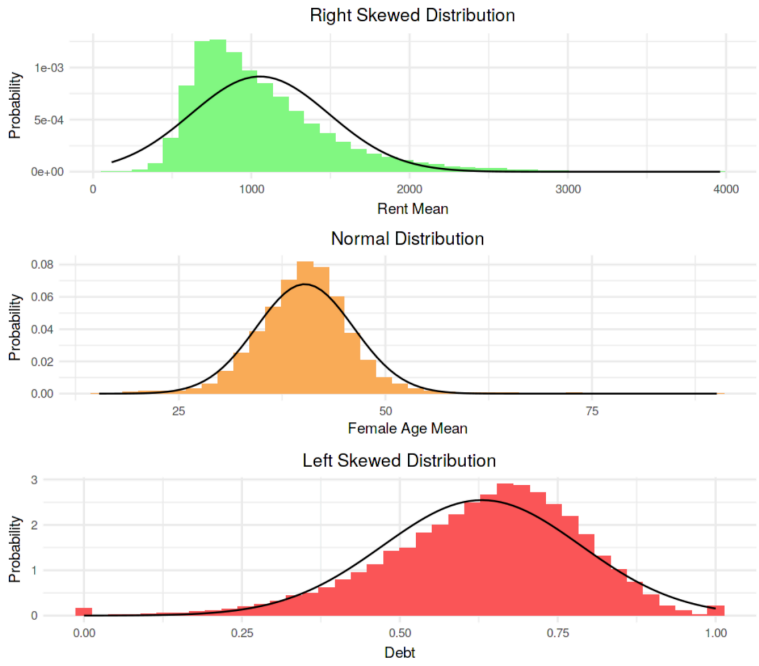
- Other data mining case
  This part focus on the data mining work by different data set. The different model will be compared and evaluated at here, then collect methods which are suitable for this data set .

- literature review Summary

2.1 Previous work in this data set

According to *Statistical analysis or a frequentist approach*[1], author main use R to do the statistical data analysis. At first, he summary each attribute by R. For numeric attribute, its mean, median, max and min value would be calculate, and represent by R shell. For String value such as state, its frequency will be summarized. For example, the New York state occur 2565 times.

After that author select all the numeric attributes using for data distribution work. The distribution result has three types, normal distribution, right skewed distribution and left skewed distribution, show in picture 2.1 below. Left skewed distribution also called as negatively-skewed[3], which has long left tail and its mean and median value also in the left. In contrast, the mean and median of right skewed distribution(positively distribution) is in right side.



Picture 2.1

Followed by, this author also use this data set to drew boxplot chart, bar chart and distribution. All the work that this author did is in the exploratory data analysis(EDA) level, which means it not reach to data mining level. These analyzing charts are benefited to help readers have better understanding of data. However, in our project, its need put data analyzing into higher level. In data mining level, we should consider how to classify or cluster data. As a result, the rest details are unnecessary for this project.

In *Insightful & Vast USA Statistics statistic EDA & EFA*[4], the author, who is also origin data provider, also did exploratory data analysis(EDA) by this data set.

The rest of work which use this data set stay in EDA level. Even EDA provide good visualization and is useful for understanding data feature.

2.2 Other data mining cases

According to *Decision Tree of Behavioral Model for Income of Indian Adults at USA*[11], the author use attributes including age, work class, education martial-status, occupation, relationship, how many hours they work per week and class, to build a decision tree model, to classify each person salary. The conclusion show in picture 2.2 below.

S.No.	Occupation	Condition	Class	Remark
1	Prof-specialty	hours-per-week<=37	<=50k	Otherwise >50k
2	Tech-support	Age<=38	<=50k	Otherwise >50k
3	Exec-managerial		>50k	
4	transport-moving		>50k	
5.	Other-service	Age<=38	<=50k	Otherwise >50k

Picture 2.2

It is known that decision tree is one of the most popular classification method[9], because it is more interpretable. Comparing with my data set, the decision tree could be use at here, because this project also use attribute value to predict or classify income, even the attribute is a little different, but basically the method is same. In my project, the income also can be predict by using finance and citizen detail such as, age, education condition, mortgage and rent etc.

In *Application of k-Means Clustering algorithm for prediction of Students' Academic Performance*[8], the author team use student academic record and k-means clustering algorithm built a model to predict student GPA. Clustering algorithm are usually used into unsupervised learn, data be cluster by their similarity and data would not be labeled[12]. For this project, the cluster also could be used for cluster data.

According to Income prediction via support vector machine[6]. The author use support vector machine to predict personal income. Support vector machine(SVM) is one of the most popular classification method. However, at here there are almost 15% error rate in classification, because multiple similar instance have different decision value.

2.3 literature review Summary

In conclusion, there are many methods that can be used in this project including both supervised and unsupervised learning. Although previous work on this data set provide lots of data information, there almost no meaningful reference for machine learn. By search other machine learning research case, the data mining method is illuminating, even the attributes are not totally similar to our data set. The decision tree, k-means and support vector machine can be use in our project. These algorithm will be tried in design algorithm step.

3 Data preprocessing

This part mainly focus on data explanation and data preprocessing.

3.1 Data explanation

This data set has 80 attributes includes 80 attributes, which mainly about finance condition, education condition and location details. In finance part, the attributes are about rent, income, mortgage and debt in geographic location. In citizen part, it contains age, population, marriage and the number of citizen who has high school degree. Explain them one by one is costly. However, if it is necessary to read all attributes explanations, the all descriptions are provide in appendix.

3.2 Data preprocessing

In order to mine the value of data, the first thing need to be completed is data preprocessing, include data cleaning, data reduction and data transformation. The reason why the data integration is not be considered at here is due to all the analyzing data come from one comma-separated value(CSV) format file. The process of combining data bases from several sources to eliminate redundancy and achieve a broader view. By the definition of data integration, combining data form different sources to reduce redundancy and make a broader view[5], only one data source cannot be integration. As a result, these three part will be the all content in the part of data preprocessing.

Firstly, read data form CSV file by python. At here we use pandas and csv library to extract data. However, the decoding error occurred in the 3196^{th} row. This is because the wrong data format, the file cannot be decode. As a result, I delete row containing error. Then it works.

3.2.1 Data Reduction

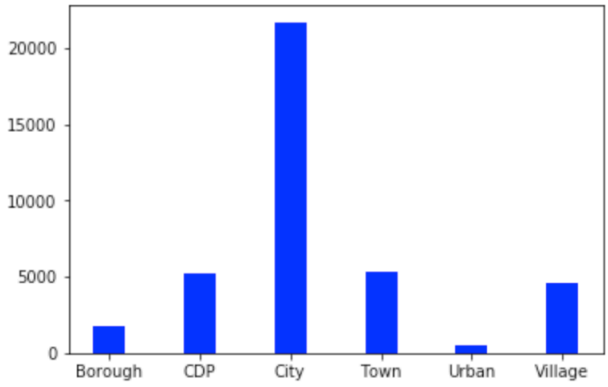
Our target is to find which attribute is related to family income. For this data set there are 39030 rows and 80 attributes, which is too much for us. Because there are some attributes may not be related to our target attribute. As a result the first thing need to do is detect and remove these attributes form our train set. In this situation, forward elimination seem more suitable.

Because our target attribute is family mean income, it is reasonable to pick up some attributes related to geographic location finance condition, such as rent, household income, family income, monthly mortgage, monthly cost and debt. As we know, the education degree also related to personal income, so I also pick up the percentage of how many people have completed high school degree in this location.

In this data, the rent is represent gross rent. Gross rent is the contract rent plus the estimated average monthly cost of utilities (electricity, gas, and water and sewer) and fuels (oil, coal, kerosene, wood, etc.) if these are paid by the renter (or paid for the renter by someone else). Gross rent is intended to eliminate differentials that result from varying practices with respect to the inclusion of utilities and fuels as part of the rental payment. The estimated costs of water and sewer, and fuels are reported on a 12-month basis but are converted to monthly figures for the tabulations.

Geographic location detail also import for our model. Each state has their own policy in American. Area's type such as city and town also import for personal income. Usually the economic situation is better than town, so the city's resident tend to have higher salary than rural area. The population situation also can reflect economic situation form side.

For the type of area, the city is the majority of instance showing picture 3.1 below. Moreover other type's Characteristics is blurred. In order reduce the interference, I only use the city type areas.



Picture 3.1

The reason why I do not consider the areas city such as New York is because there are 8172 different cites, if consider it as category attribute and consider 8172 different situations, it is unrealistic.

Another attribute may have influence to income is age situation. Actually the needed attribute is mean age in the area. However, the data set only contain female and male mean average age. Fortunately, it also provide the quantity of male and female sample. The mean age of the area is able to be calculated.

Summary above, the attribute I picked showing table below.

state	average household income
percentage of home have debt	average family income
Percentage of have high school degree	average mortgage

population	average female age
rent mean	average male age
average monthly cost	male sample quantity
female sample quantity	

3.2.2 data cleaning

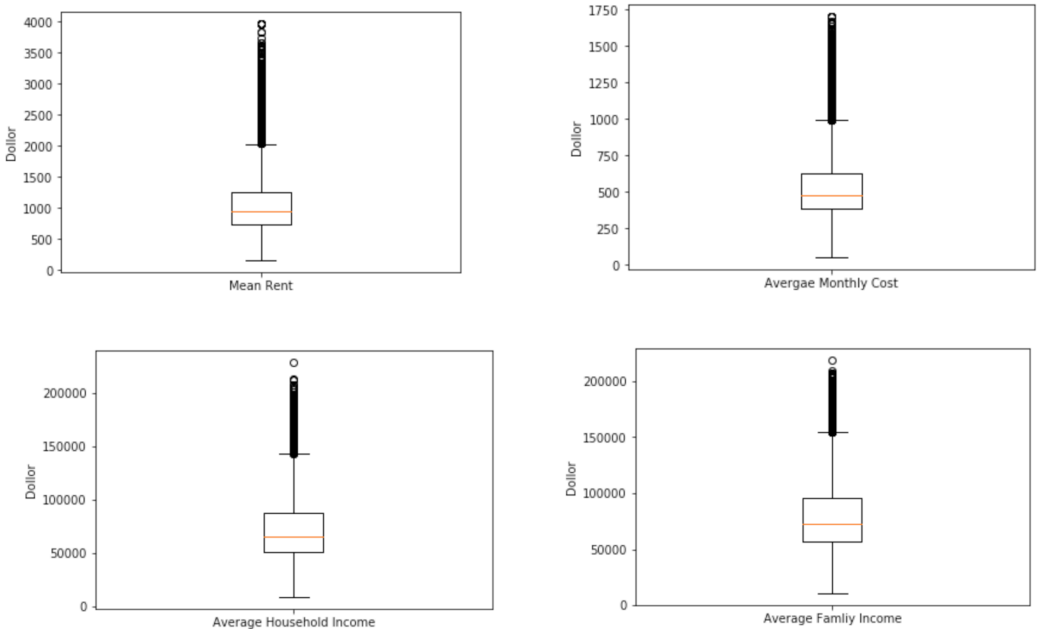
Firstly have a overlook of data set. The for each attribute its quantity of missing value show in picture 3.2 below. There are 654 rows contain missing value. Comparing with whole data set, these rows just a very little part. As a result, I remove these instances directly

```
data.isnull().sum()
```

```
STATEID              0
pop                  0
rent_mean          285
hi_mean            239
family_mean        264
hc_mortgage_mean   477
hc_mean            543
debt               391
hs_degree          162
mean_age           184
dtype: int64
```

Picture 3.2

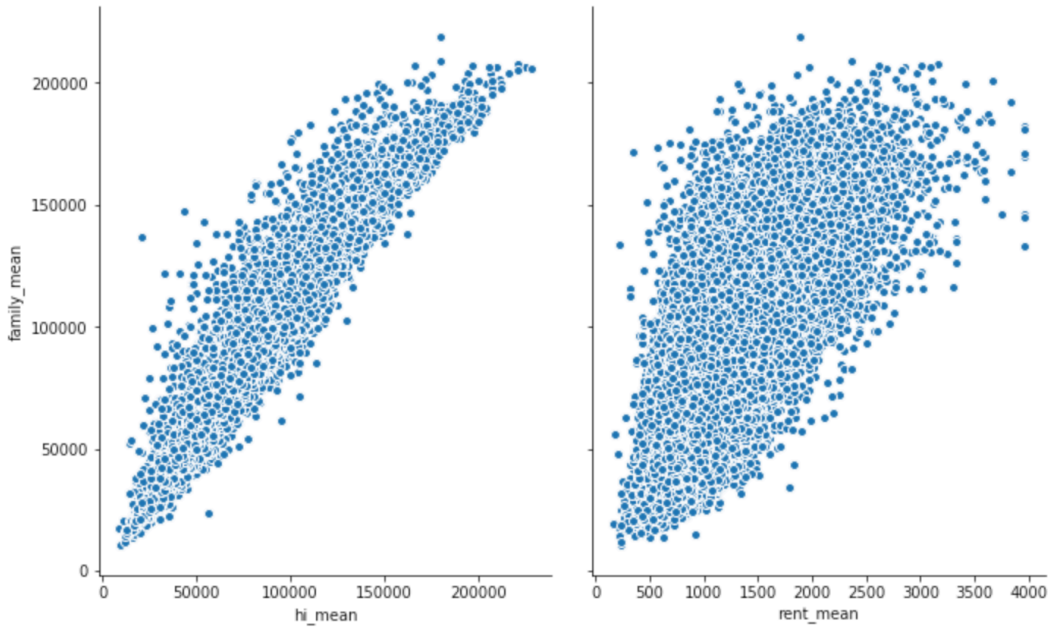
After that, I draw the boxplot chart to overview of rent, average family income, average monthly cost and average household income, showing picture 3.3 below.



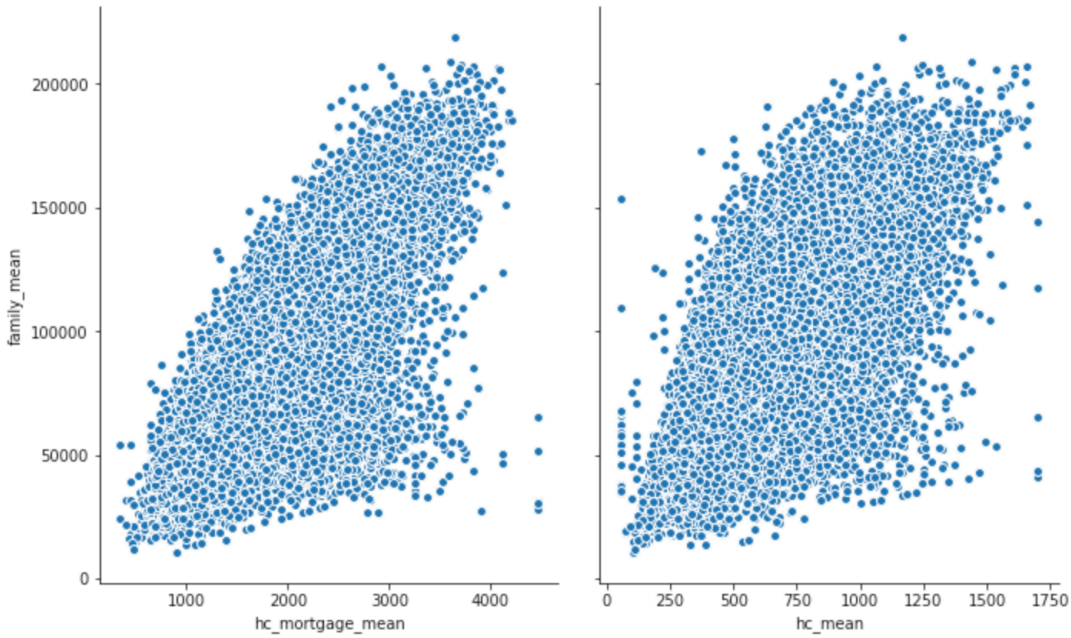
Picture 3.3

As we can see, all these attribute contain the value obvious higher than majority parts. However, these attribute should not be removed, because there are some area which belong to the rich. It is unreasonable to consider these data as outlier.

After remove city attribute, next work need to do is compare target attribute with attribute in US dollars, the picture show in picture 3.4 and 3.5 below.

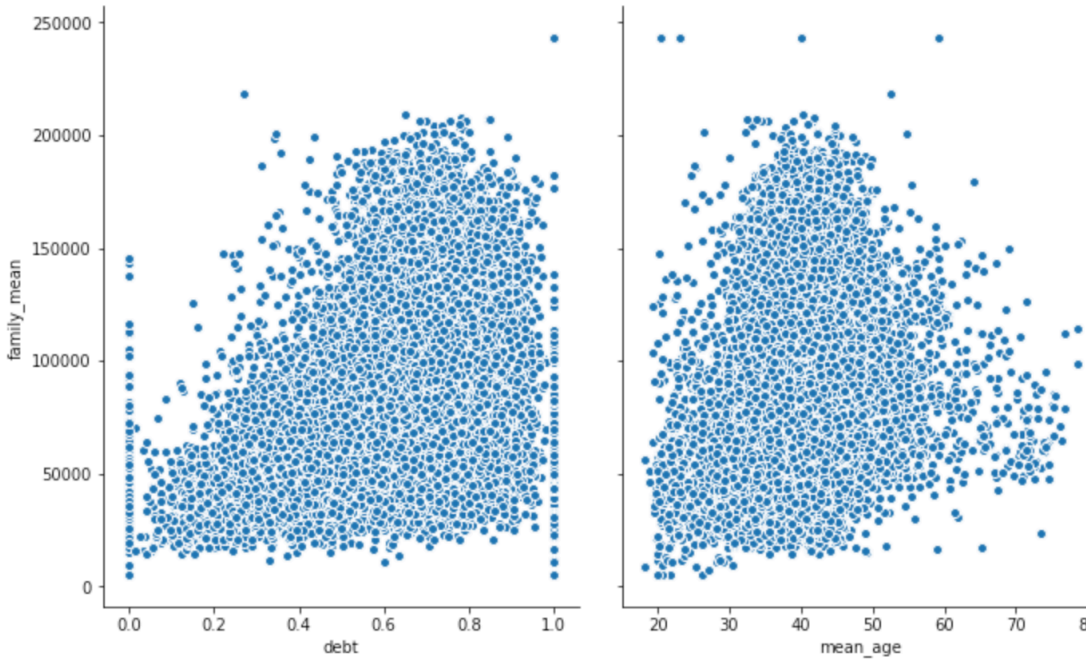


Picture3.4

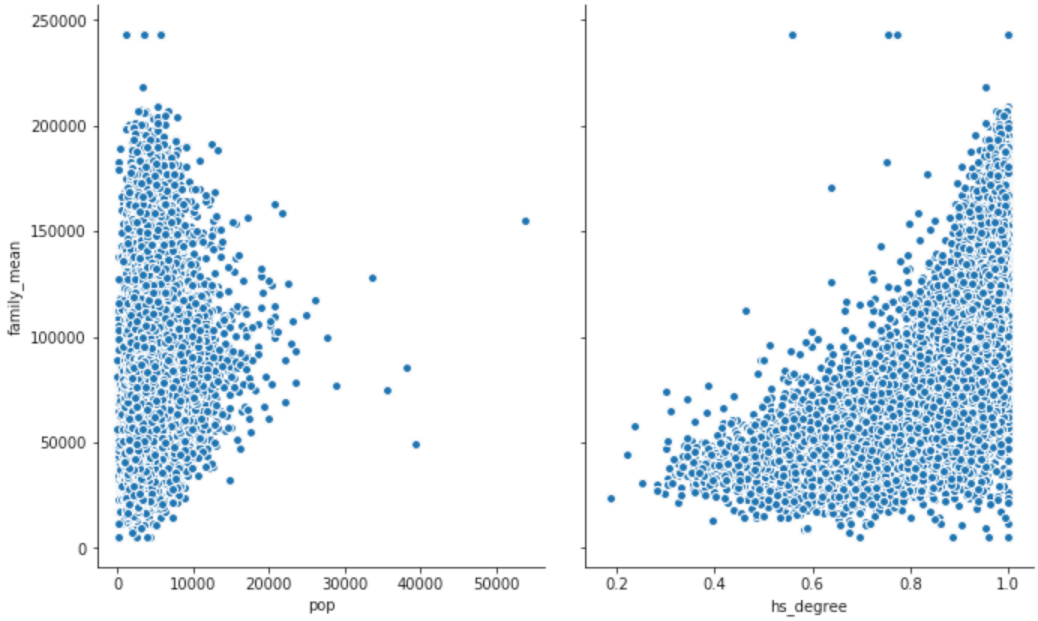


Picture 3.5

Then scatter by average family income and attributes not in US dollar show in picture 3.6 and 3.7 below.



Picture 3.6



Picture 3.7

Form these scatter charts, there is no obvious relationship between average family income type and other attributes.

It seem that we can move to next part in data preprocessing.

3.2.3 data transformation

The geographic location's average age calculate by formula below.

$$\frac{average\ female\ age \times female\ samples + average\ male\ age\ \times male\ sample}{male\ samples + female\ samples}$$

The different attributes have different unit. The age attribute calculate by year, the average household income calculate by dollar. There are huge value gap among these attribute. Because of different value scale, when put these data into model, it may lead inaccurate prediction. In order to make all attributes in a same scale, for all attribute I apply the formula below. At here, the scale is between 0 to 100

$$\frac{instance\ value - min\ value}{max\ attribute\ value - min\ attribute\ value} \times 100$$

For example, in a instance the household income is 76752, and the maximum and minimum household income is 297142 and 8858. The 76752 will be transfer to

$$\frac{76752 - 8858}{297142 - 8858} \times 100 = 23.55$$

After this operation, apart from state ID all attributes' value will between 0 to 100, showing picture 3.5 below.

	STATEID	pop	rent_mean	hi_mean	family_mean	hc_mortgage_mean	hc_mean	debt	hs_degree	mean_age
count	21063.000000	21063.000000	21063.000000	21063.000000	21063.000000	21063.000000	21063.000000	21063.000000	21063.000000	21063.000000
mean	23.294687	11.589343	23.664898	28.716396	32.597179	33.492056	30.344468	65.064740	81.405288	34.503176
std	14.931324	5.491141	12.036856	13.476550	15.131360	16.263896	14.465518	14.792992	14.593058	9.167487
min	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	12.000000	7.823031	14.707148	18.946558	21.647870	20.670007	20.199543	56.044321	74.615395	28.997931
50%	21.000000	10.867798	21.132990	25.708468	29.516265	29.103330	25.923953	66.874357	85.345686	34.297951
75%	36.000000	14.485589	29.760961	35.889068	41.078126	44.002279	36.125174	75.733408	91.999292	39.134158
max	56.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000

Picture 3.5

Form the picture above, we can split average family income to 2 classes.

Class	Value	Represent by
Low	0 to 29.52	1
High	29.52 to 100	2

Then average income attribute drop from data set's columns. After that, the data set become to, showing picture 3.6 below.

	STATEID	pop	rent_mean	hi_mean	hc_mortgage_mean	hc_mean	debt	hs_degree	mean_age	famliyType
0	2.0	12.013847	30.545606	44.844137	46.603431	47.800610	73.920268	77.514677	28.812918	2
1	2.0	9.674543	56.802675	58.111718	51.925350	40.006210	76.951515	92.255478	32.210255	2
3	2.0	4.990690	19.243478	26.365156	47.176423	26.616939	79.571534	85.944543	29.366674	2
4	2.0	15.585744	30.722129	30.898879	37.800236	38.152149	68.183848	93.907904	26.143374	2
5	2.0	14.261362	30.145074	33.231367	41.600359	37.297724	75.908907	84.197767	20.712566	2
6	2.0	15.354961	21.360270	25.626467	36.936228	27.795050	62.907463	81.755163	27.433796	2
7	2.0	19.519551	23.472919	25.841152	35.324981	41.174742	70.478831	83.821679	20.851750	2
8	2.0	14.326926	17.660212	19.887064	38.670365	30.916515	41.639926	73.371160	20.445816	1
9	2.0	8.358029	20.526926	28.002250	33.522672	35.244283	77.415704	80.600692	37.428846	2
10	2.0	8.858934	23.686390	22.501569	30.662648	44.649020	64.175416	85.345686	37.035529	2

Picture 3.6

4 Experimental design

This Section has three main parts. They are benchmark model, algorithm-1 and alogorithm-2.Support Vector Machine(SVM) is algorithm-1 and Multilayer Perceptron(MLP) is algorithm-2. The content of benchmark model will focus on how to analysis algorithm performance..

In algorithm parts (SVM and MLP algorithms), it focus on how to apply these algorithm to our data set. Both SVM and MLP. There will be briefly explanation for these two algorithm in the algorithm part.To run algorithm and model, our basic development environment is Python. I import scikit-learn(sklearn) SVM package. Sklearn is a very powerful machine learning library provided by Python third parties, which covers everything from data preprocessing to training models. Another library used here is pandas. Pandas is very useful for data operation. It can change data to data frame, which is easily to extract and modify data.

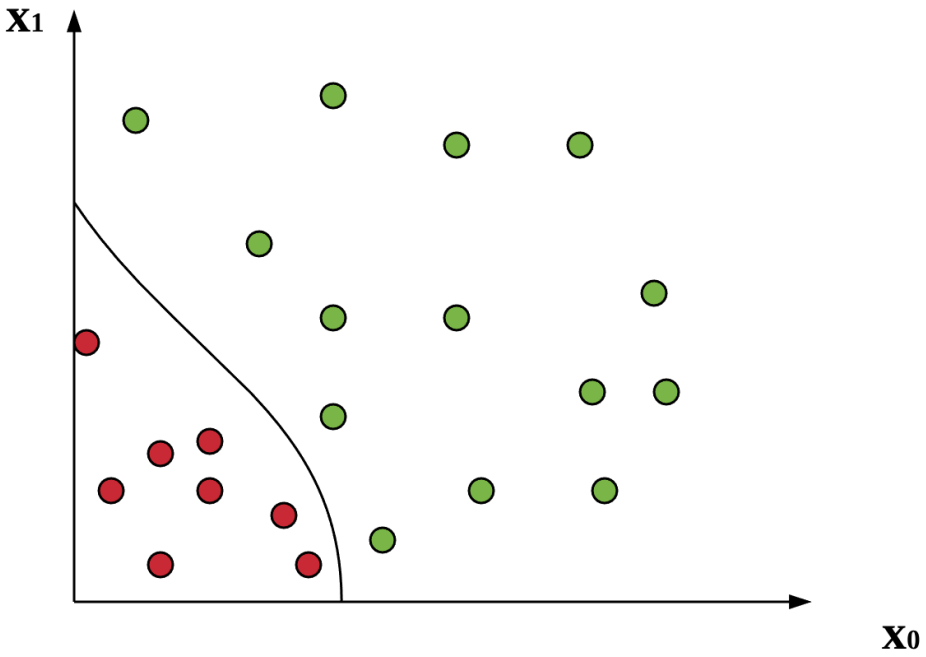
4.1 Benchmark Model

In this part, it will focus on how apply this data set to classification algorithm. I am going to use insightful & Vast USA Statistics[3] data set to check efficiency and accuracy of the classification algorithm. There 21063 row and 10 columns in our formatted data set. We only use 9 columns.

For the efficiency of the algorithm model, we mainly use time as the standard of evaluation.

- Compare the time to build a model with the same data set.
- Compare the time cost using the same test set after the model is built.

In this project, both algorithm I used are based on classification algorithm. Classification is a very important part of machine learning. Its goal is to determine sample class by certain characteristics of known samples. Classification is one of the supervised learn way. The red cycles and the green cycles represent different class instance, showing picture 4.1 below.



Picture 4.1

Assume, the red cycle represent female and the green cycle represent male. For each instance(individual), it has some characteristics(attributes) such as weight, height and foot size etc. The table below as an explanation example.

Person	X			Y
Name	Height	Weight	Foot Size	Gender
Austin	183 cm	80 kg	42	Male
Sandy	160 cm	50 kg	36	Female
Eric	177 cm	71 kg	40	Male

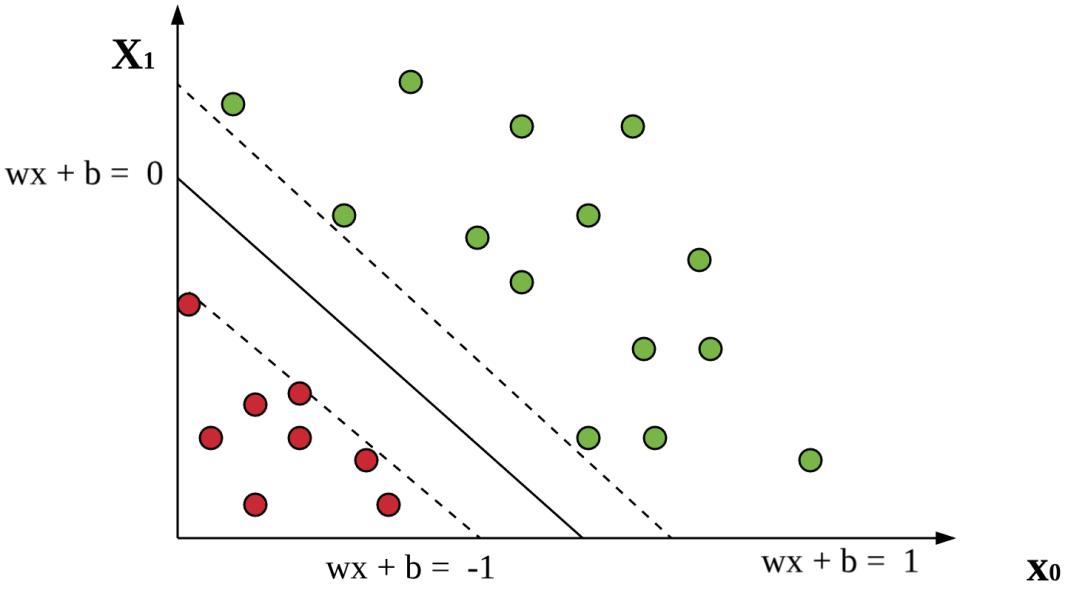
For each instance in this data set, the height, weight and foot size is the attribute set X(X = [X0(Height), X1(Weight), X2(Foot Size)]). Basically, after collect sufficient data and choose suitable classification algorithm, it is able to predict Y by giving X.

The target of classification method is to determine the best hyperplane, which is the line in picture 4.1.

4.2 Algorithm 1 - Support vector machine

SVM is a supervised learning model and related learning algorithm for analyzing data in classification and regression analysis. Generally speaking, it is a two-class classification model. The basic model is defined as the linear classifier with the largest gap in the feature space. The learning strategy is to maximize the gap and finally transform into a solution to a convex quadratic programming problem.

Basically, SVM work by support vector, for each support vector it has already classified, in train data set. The more exploration will be provided in picture 4.2 below.



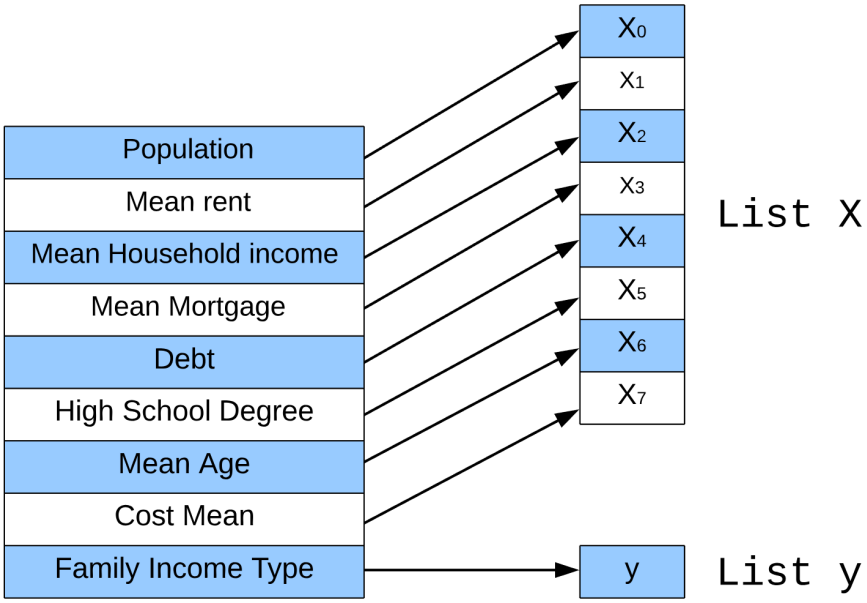
Picture 4.2

From the picture 4.2, there are two different colour cycles. The red cycles represent negative class and the green represent positive. Every cycle has two attributes, X_0 and X_1 in X set. After provide sufficient data, SVM can determine the hyperplane which has largest distance between distance.

In order to apply data set to algorithm, the first process is split data set X collection and target attribute Y. At here, the way to split data set show in below.

X		y
Population	Debt	Family type
Mean rent	High school degree	
Mean household income	Mean age	
Mean mortgage	Cost mean	

The first thing I do, is split data to X and y. For each instance, extract attributes in X and put it into X list. Also the family income type put into Y list. The process detail show in picture 4.3 below.



Picture 4.3

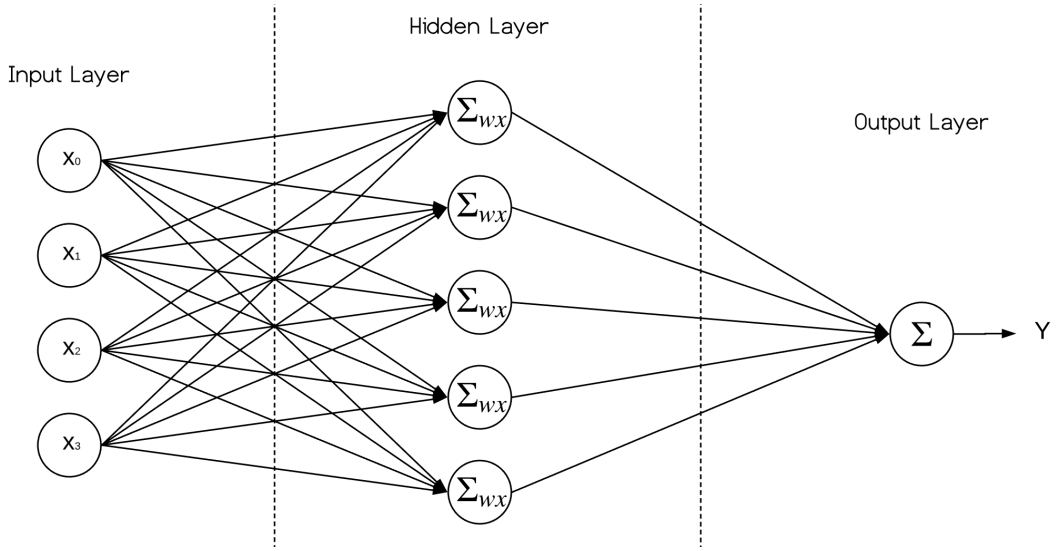
After that, the data also need be split to two parts, training data set and test data set. At here I use 10% data as the test data set. There are 21063 data, so there should be 2106 data in the test data set. The way I collect test data set is generate 2106 not-repeating values which are between 0 to 21063. These value represent the index in data set. Then split the data whose index in these numbers as test set, and the rest as the training data set.

Then build the model by the train data set. Initial the Support Vector Classification model, which provide by sklearn, fit the X and Y of the training data set. After fit all training data, the model is built. Now the it can predict the family in come type by the X.

Now put the X of test data set into model to predict family income type. Then the model will return a list of family type. Next step is to compare the prediction result with Y of test data set to calculate the right rate.

4.3 Algorithm 2 - Multi-Layer Perceptron

Multi-Layer Perceptron(MLP), is also a classification algorithm. The basic structure shows in picture 4.4 below.



Picture 4.4

Each circle in the above picture 4.4 is a neuron, and each line represents the connection between neurons. We can see that the above neurons are divided into multiple layers, the neurons between the layers are connected, and the neurons in the layers are not connected. The leftmost layer is called the input layer. This layer is responsible for receiving input data. The rightmost layer is called the output layer. We can get the neural network output data from this layer. The layer between the input layer and the output layer is called a hidden layer. The input layer, is each attribute in X collection. The Y is the result after operation. Through continuously alter weight by backpropagation and training set, until the weight is suitable. After that, the MLP model will be created. Same as SVM, it can get the Y by input X.

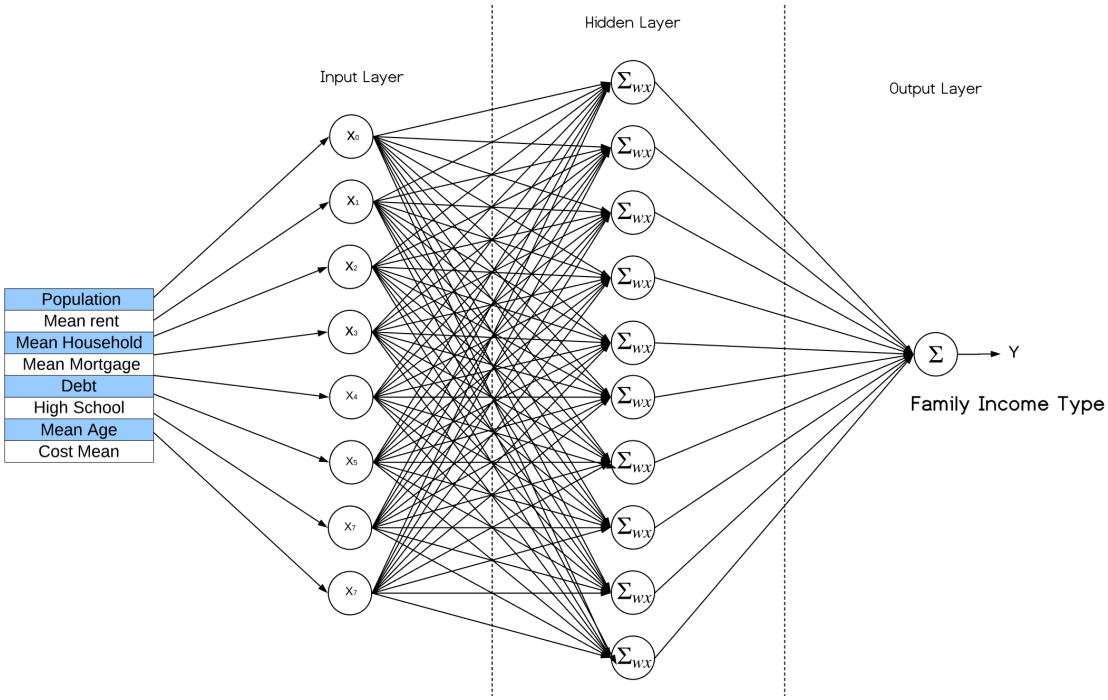
The processes of splitting data to X and Y and training data set and test data set is same to SVM.

Input Layer		Output Layer
Population	Debt	Family type
Mean rent	High school degree	
Mean household income	Mean age	
Mean mortgage	Cost mean	

Put each attribute in X to a neuron in input layer. Each line in picture 4.5 have a weight. For each neuron in hidden layer and output layer, calculate the value by formula below.

$$\sum wx$$

If the prediction is not correct, then alter the weight by backpropagation. After fitting all instances, the model is built. Then it can predict family income type by X.



Picture 4.5

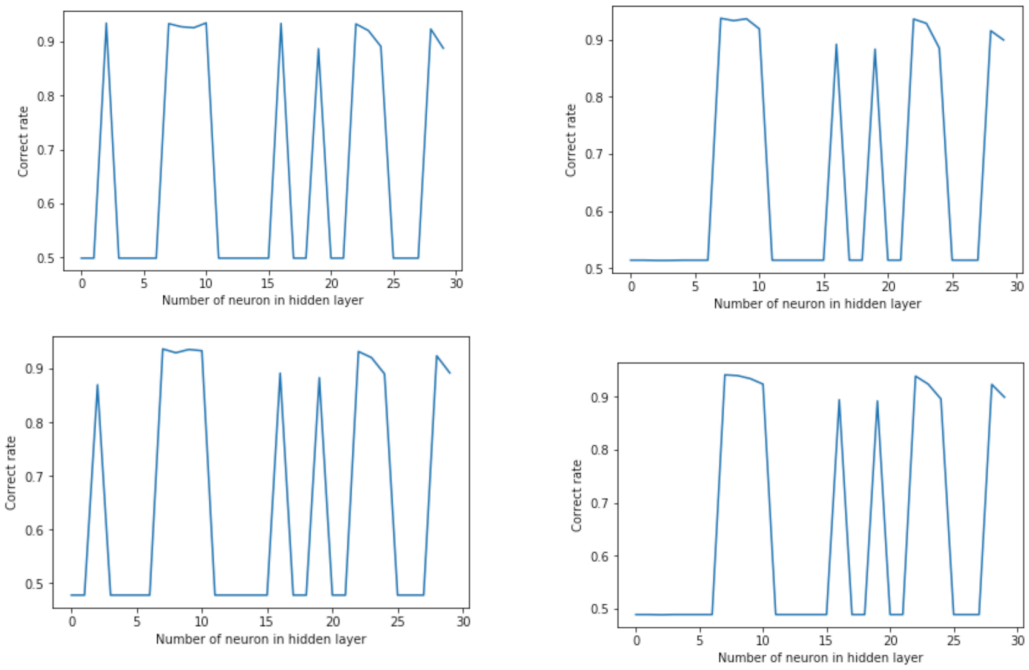
Same as SVM, the accuracy of test data set will be calculated.

The reason why choose SVM and MLP is because they all classification algorithm. It is easily to analyze the result of two algorithms.

5 Result Analysis

5.1 Accuracy

For MLP, the first thing is find the most suitable number of neuron in hidden layer. At here sample 10% data in X and Y as test data set, and the rest is the training data. Then apply the training data set to MLP 30 time, but each time with the different number of neuron. Let the number of neuron form 1 to 30 and calculate the accuracy. In order to make it accurate, I run this process 4 times. Then print the line chart, showing picture 5.1 below. Find the most suitable number of neuron whose accuracy is the highest.

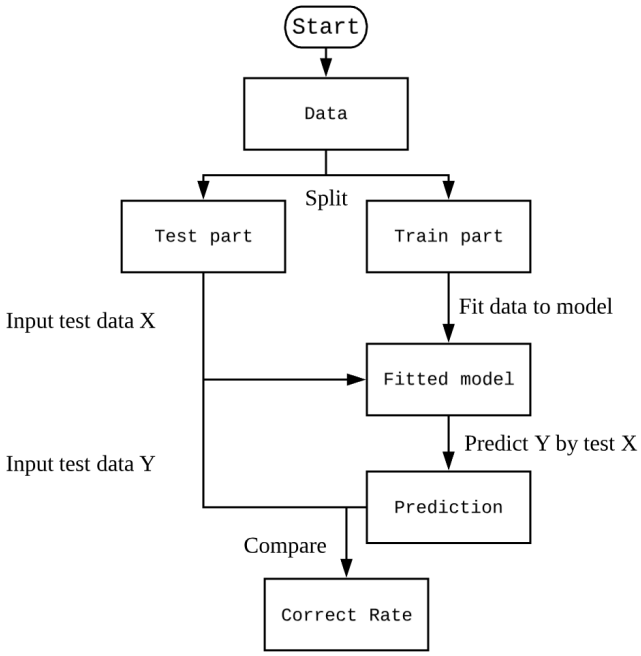


Picture 5.1

From the picture above and the program, there are 3 time that the most suitable number of neuron is 8. As a result, we de consider the most suitable number of neuron is 8 on applying these data set to MLP.

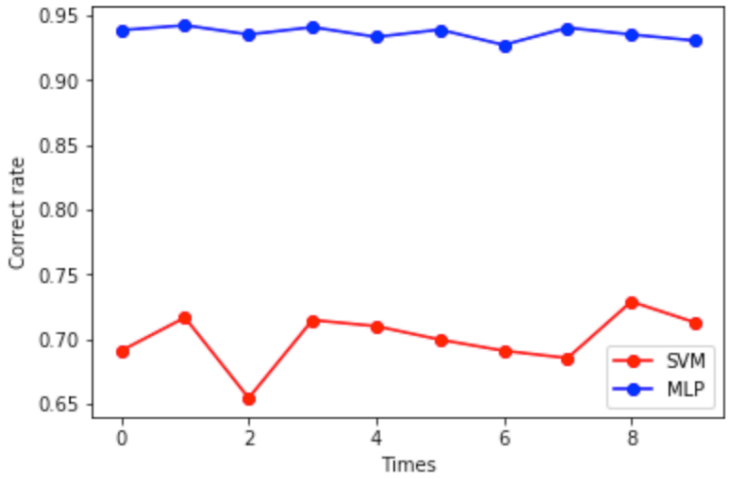
However, there is no parameter for SVM. As a result, just split data to training part and test part. Then apply training part to SVM model. Then it can predict family income type by X.

For each model, the flow chart of how to calculate the accuracy showing picture 5.2 below.



Picture 5.2

Then compare two model. The way to compare two model is to calculate the accuracy by same data set. At here I use same data set, but split different training part and test part 10 time. For each time, print the accuracy showing in picture 5.3 below

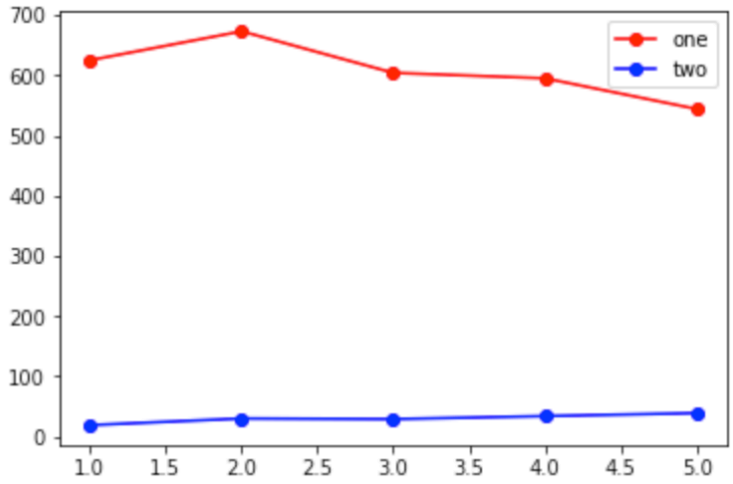


Picture 5.3

As we can see, the accuracy of SVM is obvious lower than MLP. The highest accuracy of SVM is low than 75%, in contrast the lowest accuracy of MPL is over than 91%.

For this data set using MLP is accurate enough. However, for SVM is not accurate enough. Because random predict family type, these still have around 50% accuracy.

To detect the reason why the accuracy of SVM model is not high enough. It is necessary to know the which family income is wrong classification. Count the number of wrong classification showing picture 5.4 below. The x axis represent in which time. For example. The first time, there are around 650 wrong classification which should be classified as one, but the predict value is two.



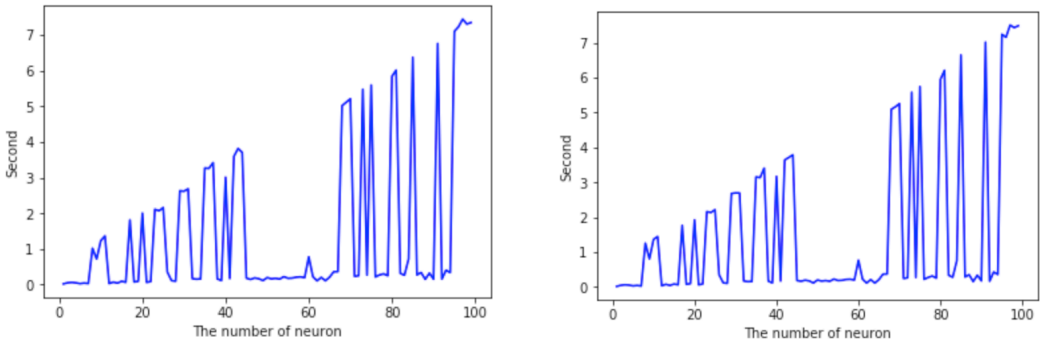
Picture 5.4

Form picture 5.4, it is easily to wrong classify the low family income type to high income type. However, the high income family type have almost correct classification.

In my view, the reason why there are so many low family income location, which has been wrong classification in SVM, is due to some attributes in data set contain outlier. It may lead wrong classification[7]. However at the same time, I feel that it unwise to directly delete those instances that have outliers. Because each instance may have value for the model, and these instances are also present

5.2 Efficiency

Firstly, find whether there is relationship between the number of neuron in hidden layer and the time cost. Let the number of neuron in hidden layer form 1 to 100. For each number record its time cost. Then print it as line chart two times, showing 5.5 below.

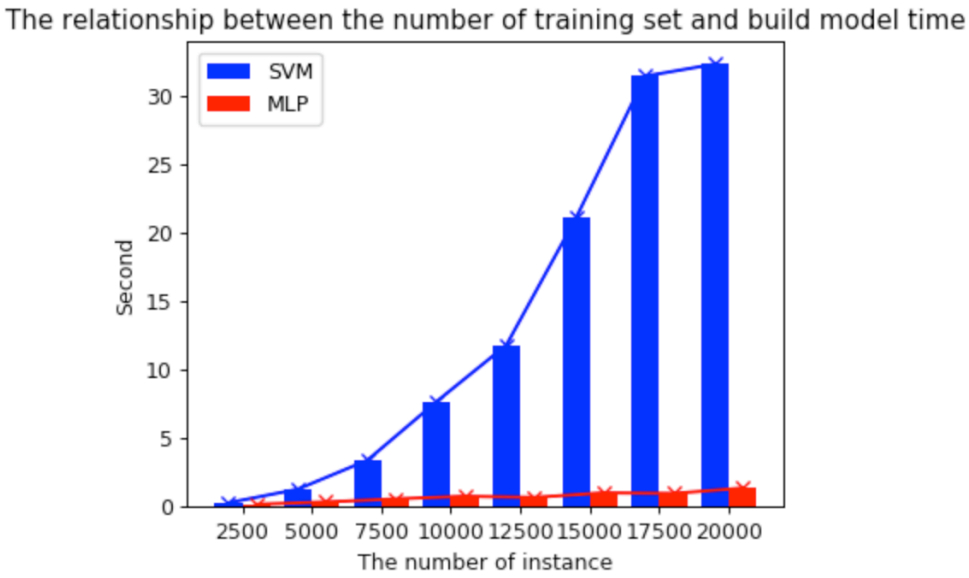


1st Time 2rd Time

Picture 5.5

Basically, the first time and second time is same. We can know that, In some cases, the number of neuron and time cost are positively correlated. In other ranges, the number of neuron and time cost have little relationship, and the time cost is relatively stable in a very low range.

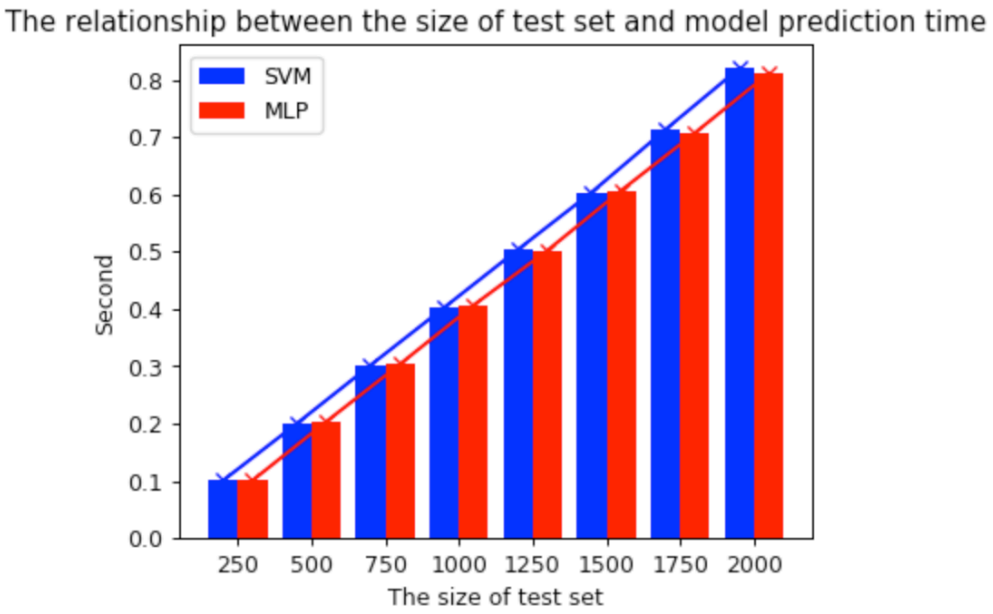
Now compare the relationship between the number of data sets used for training and the time of build model, showing in picture 5.6 below. Models are built using training sets of size 2500, 5000, 7500, 10000, 12500, 15000, 17500 and 20000 respectively.



Picture 5.6

As we can see, both for SVM and MLP, the time cost of build model and the size of training set are positively correlated. However, under the same size of training set, The growth rate of SVM is significantly higher than the growth rate of MLP. The growth rate of MLP is linear. SVM time cost is exponential growth. In any case, the time cost of MLP is lower than SVM.

After building the model with the same training set, we want to know which model is more efficient. The picture 5.7 below show this relationship.



Picture 5.7

As can be seen from the above figure, for both models, prediction time and the size of test set is positively correlated. Moreover in the case where the size of test set is equal, the prediction time of the two models are basically the same.

5.3 Analysis Summary

The summary table shows below.

Algorithm	SVM	MLP
Accuracy	Low	High
Efficiency	Low	High

As we can see both for accuracy and efficiency, MLP is better than SVM for this data set.

6 Conclusion

The main focus of this article is on classification algorithms and how to apply classification algorithms to data sets. The MLP is very good. To use MLP, the most important thing is to find the right number of neurons in the hidden layer. If the right number of neurons are not found in the hidden layer, it is very likely that the predicted values are extremely inaccurate. As a result, the most important thing for MLP is to find the right number of nerves in the hidden layer.

For SVM, I think the most important thing to improve the accuracy of SVM model prediction is in the process of data preprocessing. Accurate predictions need to remove outliers from attributes in the data set.

For these two algorithms, the internal running process is difficult to observe, just like a black box.

Under the same conditions, MLP is more efficient than SVM from modeling to prediction.

In general, I think the longest and most important part of data mining is data preprocessing. If the data preprocessing is not done well, there will be many problems with the model that is built later, such as the deviation of the predicted values. After great data preprocessing, the probability of success of the model will be much higher.

References

[1] J. Bachmann.(2018, Jul). "Statistical analysis or a frequentist approach". Available: https://www.kaggle.com/janiobachmann/statistical-analysis-a-frequentist-approach?scriptVersionId=5448479

[2] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth., "From data mining to knowledge discovery in databases", AI magazine, vol. 17, issue 3. pp. 37-54. 1996

[3] A. Geiger.(2018, May). "insightful & Vast USA Statistics". Available: https://www.kaggle.com/goldenoakresearch/us-acs-mortgage-equity-loans-rent-statistics#real_estate_db.csv

[4] A. Geiger.(2018, Jun). "Insightful & Vast USA Statistics EDA & EFA". Available: https://www.kaggle.com/alexgeiger/insightful-vast-usa-statistics-eda-efa?scriptVersionId=3740787

[5] J. Han, M. Kamber and J. Pei., Data mining concepts and techniques, 3rd ed. Massachusetts. USA: Elsevier, 2012

[6] A. Lazar.(2012). "Income prediction via support vector machine". Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.387.5068&rep=rep1&type=pdf

[7] W. Mike and L. Cheung.(2010, Oct). "outlier and influence diagnostics for meta-analysis". Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jrsm.11

[8] O. Oyelade, O. Oladipupo and I. Obagbuwa.(2010). "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance". Available: https://arxiv.org/pdf/1002.2425.pdf

[9] S. Reddy, K. Achari and V. Vasu.(2013). "Decision trees for training data sets containing numerical attributes with measurement errors". International Journal of Advanced Research in Computer Science[online]. vol. 4, issue 4 Available: https://search-proquest-com.ezproxy.lib.monash.edu.au/docview/1443749061?accountid=12528

[10] Stephanie.(2018). "Skewed Distribution: Definition". Available: http://www.statisticshowto.com/probability-and-statistics/skewed-distribution/

[11] R. Soni and M. Mandot.(2012). "Decision tree of behavioral model for income of Indian adults at USA". International Journal of Advanced Research in Computer Science[online]. vol. 3, issue 3. Available: https://search-proquest-com.ezproxy.lib.monash.edu.au/docview/1443724664?accountid=12528

[12] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl.(2001). "Constrained K-means Clustering with Background Knowledge". Proceedings of the Eighteenth International Conference on Machine Learning[online] pp. 557-P584. Available: https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf

Appendix

Attribute Name	Attribute Description
UID	The ID of the location of which you are analyzing. ID location compatible across all Golden Oak Research Locations.
STATEID	The state code reported by the U.S. Census Bureau for the specified geographic location.
BLOCKID	Block ID of tract. If there is no specified block id the location is then a tract. The maximum number of blocks for any given track is 9.
STATE	The state name reported by the U.S. Census Bureau for the specified geographic location.
state_ab	The abbreviated state name reported by the U.S. Census Bureau for the specified geographic location.
country	The county name reported by the U.S. Census Bureau for the specified geographic location.
city	The closest city name reported by the U.S. Education Department by the closest school relative to the Census Location.
place	The place name reported by the U.S. Census Bureau for the specified geographic location.
type	The place Type reported by the U.S. Census Bureau for the specified geographic location.
primary	Defines whether the location is a tract location or a block group.
zip_code	The closest zip code reported by the U.S. Education Department by the closest school relative to the Census Location.
Area_code	The area code reported by the U.S. Census Bureau of the closest geographic location with area code information.
lat	The latitude of geographic location.
lng	The longitude of geographic location.
ALand	The Square area of land at the geographic or track location.
AWater	The Square area of water at the geographic or track location.
pop	Male & female population of geographic location
male_pop	Male population of geographic location.
female_pop	female population of geographic location.

second_mortgage	percent of houses with a second mortgage
home_equity	Percentage of homes with a home equity loan.
home_eqiity_second_mortgage	Percentage of homes with a second mortgage and home equity loan.
home_equity_second_mortgage	Percentage of homes with a second mortgage and home equity loan.
debt	Percentage of homes with some type of debt.
second_mortgage_cdf	Cumulative distribution value of one minus the percentage of homes with a second mortgage. The value is used as a performance feature.
home_equity_cdf	Cumulative distribution value of one minus the percentage of homes with a home equity loan. The value is used as a performance feature.
debt_cdf	Cumulative distribution value of one minus the percentage of homes with any home related debt. The value is used as a performance feature.
hs_degree	Percentage of people with at least high school degree.
hs_degree_male	Percentage of males with at least high school degree.
hs_degree_female	Percentage of females with at least high school degree.
hc_mortgage_mean	The mean Monthly Mortgage and Owner Costs of specified geographic location.
hc_mortgage_median	The median Monthly Mortgage and Owner Costs of the specified geographic location.
hc_mortgage_stdev	The standard deviation of the Monthly Mortgage and Owner Costs for a specified geographic location.
hc_mean	The number of samples used in the statistical calculations
hc_median	The median Monthly Owner Costs of a specified geographic location.
hc_stdev	The standard deviation of the Monthly Owner Costs of a specified geographic location.
hc_samples	The samples used in the calculation of the Monthly Owner Costs statistics.
hc_sample_weight	The samples used in the calculation of the Monthly Owner Costs statistics.
rent_mean	The mean gross rent of the specified geographic location.
rent_median	The mean gross rent of the specified geographic location.

rent_stdev	The standard deviation of the gross rent for the specified geographic location.
rent_samples	The number of gross rent records used in the statistical calculations
rent_sample_weight	The sum of gross rent weight used in calculations.
rent_gt_10 (CDF)	The empirical distribution value that an individual's rent will be greater than 10% of their household income in the past 12 months.
rent_gt_15 (CDF)	The empirical distribution value that an individual's rent will be greater than 15% of their household income in the past 12 months.
rent_gt_20 (CDF)	The empirical distribution value that an individual's rent will be greater than 20% of their household income in the past 12 months.
rent_gt_25 (CDF)	The empirical distribution value that an individual's rent will be greater than 25% of their household income in the past 12 months.
rent_gt_30 (CDF)	The empirical distribution value that an individual's rent will be greater than 30% of their household income in the past 12 months.
rent_gt_35 (CDF)	The empirical distribution value that an individual's rent will be greater than 35% of their household income in the past 12 months.
rent_gt_40 (CDF)	The empirical distribution value that an individual's rent will be greater than 40% of their household income in the past 12 months.
rent_gt_50 (CDF)	The empirical distribution value that an individual's rent will be greater than 50% of their household income in the past 12 months.
rent_universe_samples (CDF)	The size of the renter-occupied housing units sampled universe for the calculations.
rent_used_samples	The number of samples used in the household income by gross rent as percentage of income in the past 12 months calculation
family_income_mean	The mean family income of the specified geographic location.
family_income_median	The median family income of the specified geographic location.
family_income_stdev	The standard deviation of the family income for the specified geographic location.
family_income_families	The number of families used in the statistical calculations

hi_mean	The mean household income of the specified geographic location.
hi_median	The median household income of the specified geographic location.
Hi_stdev	The standard deviation of the household income for the specified geographic location.
hi_samples	The number of households used in the statistical calculations
hi_sample_weight	The number of households weighted used in the statistical calculations
male_age_mean	The mean male age of the specified geographic location.
male_age_median	The median male age of the specified geographic location.
male_age_stdev	The standard deviation of the male age for the specified geographic location.
male_age_sample_weight	The number of male age weighted used in the statistical calculations
male_age_samples	The number of male age used in the statistical calculations
female_age_mean	The mean female age of the specified geographic location.
female_age_median	The median female age of the specified geographic location.
female_age_stdev	The standard deviation of the female age for the specified geographic location
female_age_sample_weight	The number of female age weighted used in the statistical calculations
female_age_samples	The number of female age used in the statistical calculations