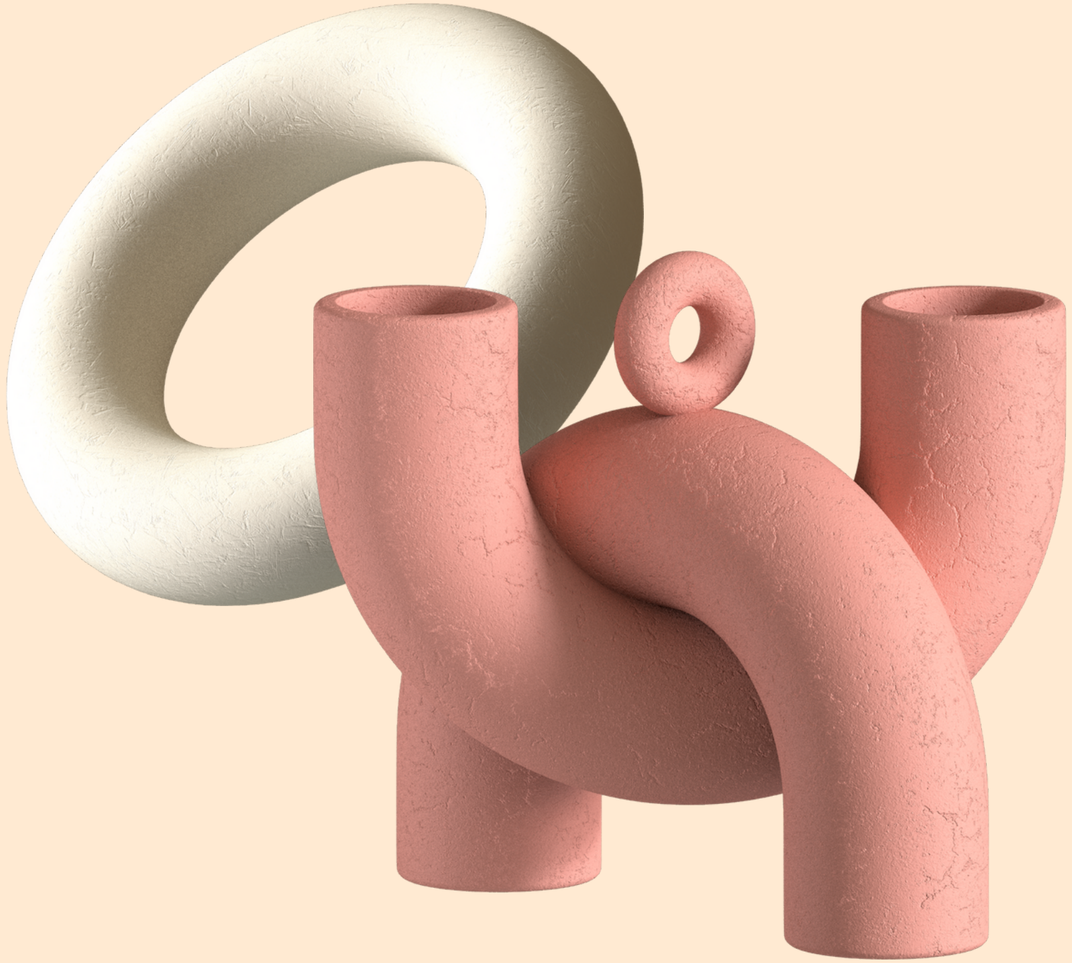
WIE3007 Data Mining & Warehousing

Group Project

Income Group Modeling



Group Member:

Hong Jia Herng (U2005313/1)
Lee Hui Xin (U2005353/1)
Chai Nam Chi (U2005421/1)
Cheong Hui Ting (U2005292/1)
Liewn Wan Chyi (U2005418/1)

Income Group Modeling

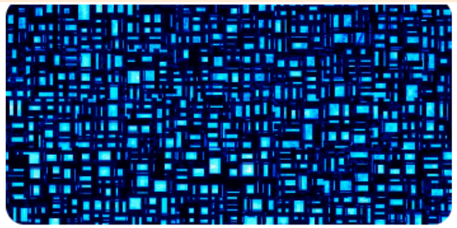
What to Know

1.0 Dataset Introduction & Summary

Chosen Dataset: Adult Data Set (Cencus Income dataset)

Adult Data Set (Census Income dataset)

Predict whether income exceeds $50K/yr based on census data



Data Card | Code (2) | Discussion (1)

About Dataset

The dataset used is US Census data which is an extraction of the 1994 census data which was donated to the UC Irvine's Machine Learning Repository.

The data contains approximately 32,000 observations with over 15 variables.

The dataset was downloaded from:

http://archive.ics.uci.edu/ml/datasets/Adult.

The dependent variable in our analysis will be income level and who earns above $50,000 a year using SQL queries, Proportion Analysis using bar charts and Simple Decision Tree to understand the important variables and their influence on prediction.

Usability
7.65

License
Unknown

Expected update frequency
Never

Tags

Business | Social Science | Beginner

32, 561 instances
15 columns

Link to dataset: https://www.kaggle.com/datasets/kritidoneria/adultdatasetxai?select=Metadata.txt.names

1.0 Dataset Introduction & Summary

Dataset Summary

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
12	_0	Num	8			0
5	_13	Num	8			13
1	_39	Num	8	BEST12.	BEST32.	
13	_40	Num	8			40
11	_2174	Num	8			2174
3	_77516	Num	8			77516
7	_Adm_clerical	Char	17			Adm-clerical
4	_Bachelors	Char	12			Bachelors
10	_Male	Char	6			Male
6	_Never_married	Char	21			Never-married
8	_Not_in_family	Char	14			Not-in-family
2	_State_gov	Char	16			State-gov
14	_United_States	Char	18			United-States
9	_White	Char	18			White
15	___50K	Char	5			<=50K

File Import Summary

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	age	Num	8	BEST12.	BEST32.
11	capital_gain	Num	8	BEST12.	BEST32.
12	capital_loss	Num	8	BEST12.	BEST32.
4	education	Char	12	$12.	$12.
5	education_num	Num	8	BEST12.	BEST32.
3	final_weight	Num	8	BEST12.	BEST32.
15	gross_income	Char	5	$5.	$5.
13	hours_per_week	Num	8	BEST12.	BEST32.
6	marital_status	Char	21	$21.	$21.
14	native_country	Char	18	$18.	$18.
7	occupation	Char	17	$17.	$17.
9	race	Char	18	$18.	$18.
8	relationship	Char	14	$14.	$14.
10	sex	Char	6	$6.	$6.
2	work_class	Char	16	$16.	$16.

Updated Variable Summary

1.0 Dataset Introduction & Summary

Dataset Summary

	id	timestamp	age	work_class	final_weight	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_wee
21261	77	2022-08-30 11:00:00	32	Private	272376	Assoc-acdm	12	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	4
27460	17	2023-05-16 04:00:00	28	Private	163772	HS-grad	9	Married-civ-spouse	Other-service	Husband	Other	Male	0	0	4
26623	44	2023-04-11 07:00:00	26	Private	39092	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	White	Male	4064	0	5
4410	73	2020-09-27 18:00:00	60	Private	181953	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	2
9286	3	2021-04-18 22:00:00	27	Private	401723	HS-grad	9	Never-married	Adm-clerical	Not-in-family	Black	Female	0	0	4

Synthesised timestamp and id column

1.0 Dataset Introduction & Summary

Column Metadata

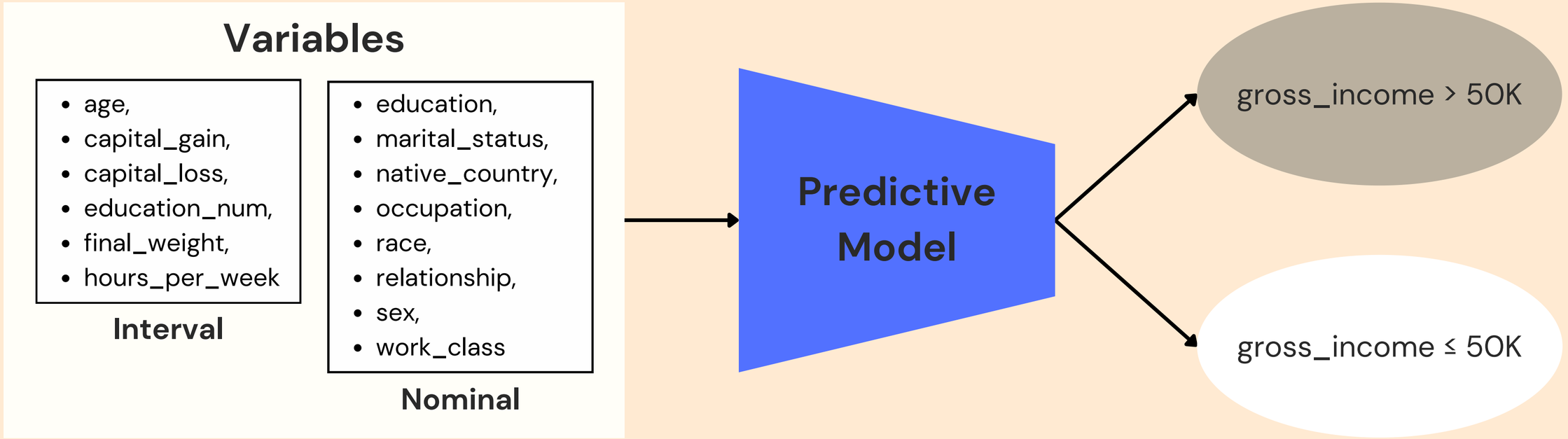
Columns	Description	Datatype
age	The age of adult	Numerical (interval)
capital_gain	The income of the adult from investment sources other than working salary	Numerical (interval)
capital_loss	The loss of adult on the investment	Numerical (interval)
education	The highest education level of the adult	Categorical (ordinal)
education_num	The numerical representation of the “education” variable	Numerical (interval)
final_weight	The number of units in the target population that the responding unit represents	Numerical (interval)
gross_income	The income group of the adult, either more tham $50,000 or less than or equal to $50,000	Categorical (ordinal)

1.0 Dataset Introduction & Summary

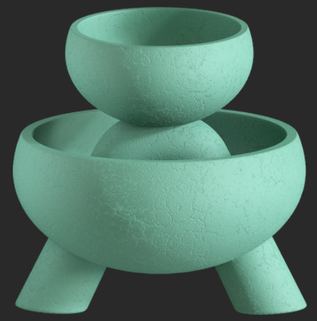
Column Metadata

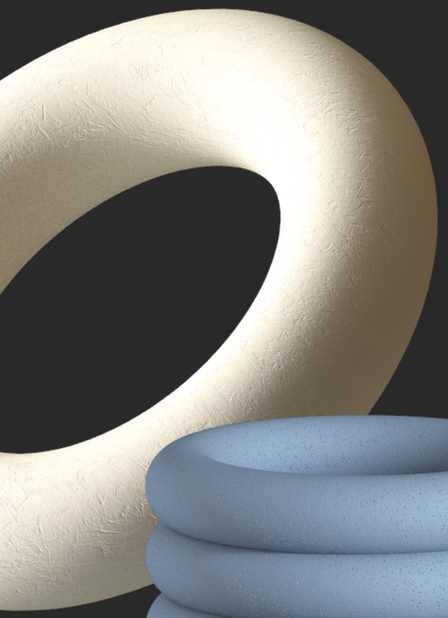
hours_per_week	The working hours of the adult per week	Numerical (interval)
marital_status	The marital status of the adult	Categorical (ordinal)
native_country	The country where the adult born in	Categorical (ordinal)
occupation	The job title of the adult	Categorical (ordinal)
race	The race of the adult	Categorical (ordinal)
relationship	The relationship status of the adult	Categorical (ordinal)
sex	The gender of the adult	Categorical (ordinal)
work_class	The company category that the adult worked at	Categorical (ordinal)

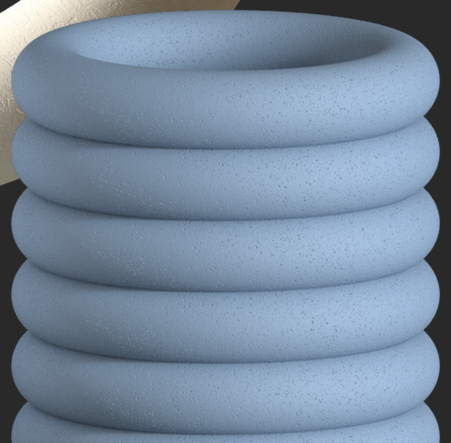
2.0 Analysis / Modeling Goal

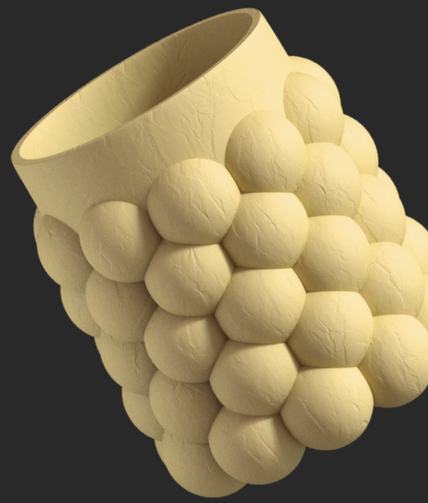


Using the input variables of an individual, to predict the gross income group of him/her.









3.0
Application of SAS SEMMA Methodology

3.1 Sample

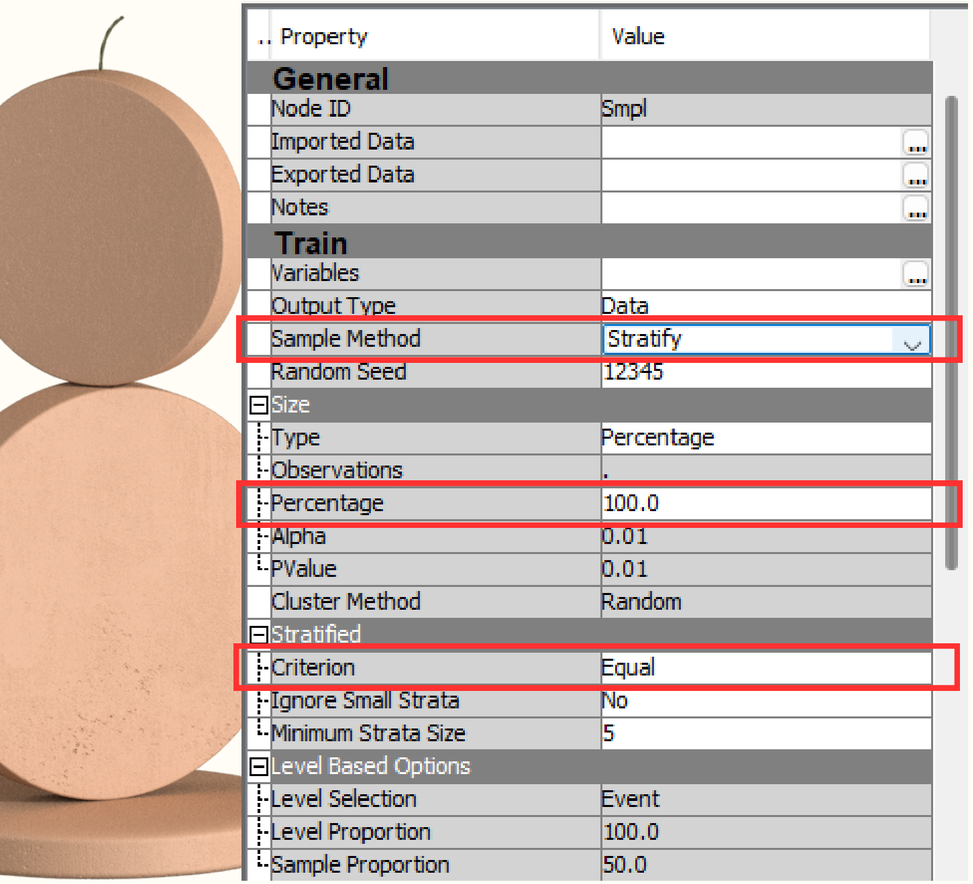
Setting the Role of the Target Variable

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
age	Input	Interval	No		No	.	.
capital_gain	Input	Interval	No		No	.	.
capital_loss	Input	Interval	No		No	.	.
education	Input	Nominal	No		No	.	.
education_num	Input	Interval	No		No	.	.
final_weight	Input	Interval	No		No	.	.
gross_income	Target	Nominal	No		No	.	.
hours_per_we	Input	Interval	No		No	.	.
marital_status	Input	Nominal	No		No	.	.
native_country	Input	Nominal	No		No	.	.
occupation	Input	Nominal	No		No	.	.
race	Input	Nominal	No		No	.	.
relationship	Input	Nominal	No		No	.	.
sex	Input	Nominal	No		No	.	.
work_class	Input	Nominal	No		No	.	.

Change the role of “gross_income” to “Target” in the File Import node.

3.1 Sample

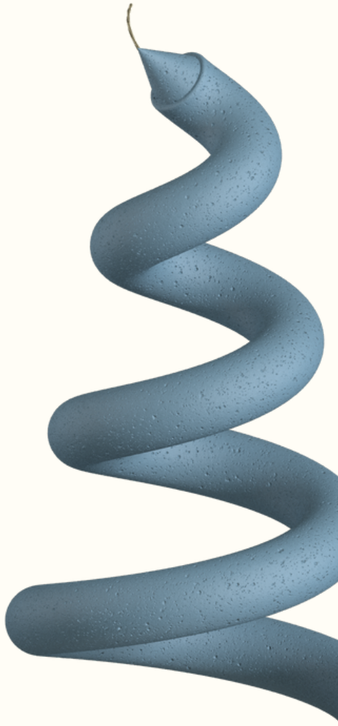
Handling Class Imbalanced Issue



Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	100.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0

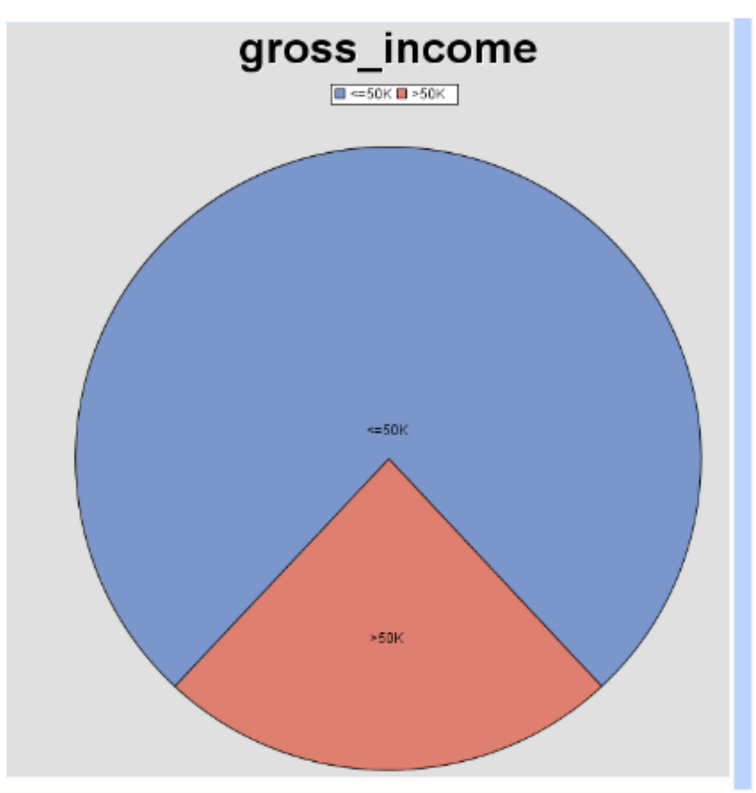
Applying stratified sampling:

- Set “Sampling Method” as “Stratify”:
  - To ensure that each strata has an equal chance of being represented in the sample.
- Set “Criterion” as “Equal”:
  - To ensure the resulting sample have same number of instances for each target class.
- Set “Percentage” as 100%:
  - To sample as many instances as possible while maintaining the equal criterion.

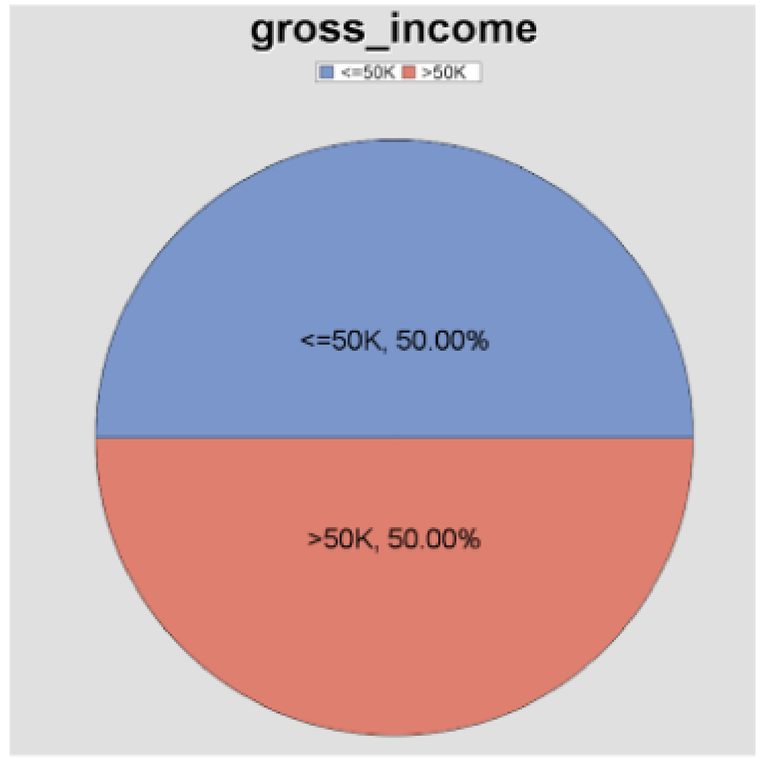


3.1 Sample

Handling Class Imbalanced Issue



Initial distribution of gross income



Distribution of gross income after sampling

3.1 Sample

Data Partitioning

.. Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Stratified
Random Seed	12345
Data Set Allocations	
Training	80.0
Validation	20.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/21/24 1:47 PM
Run ID	4783045b-583b-bd4c-b476-d
Last Error	
Last Status	Complete
Last Run Time	1/21/24 2:07 PM
Run Duration	0 Hr. 0 Min. 1.92 Sec.
Grid Host	
User-Added Node	No

Result:

Results - Node: Data Partition Diagram: ExploreData

File Edit View Window

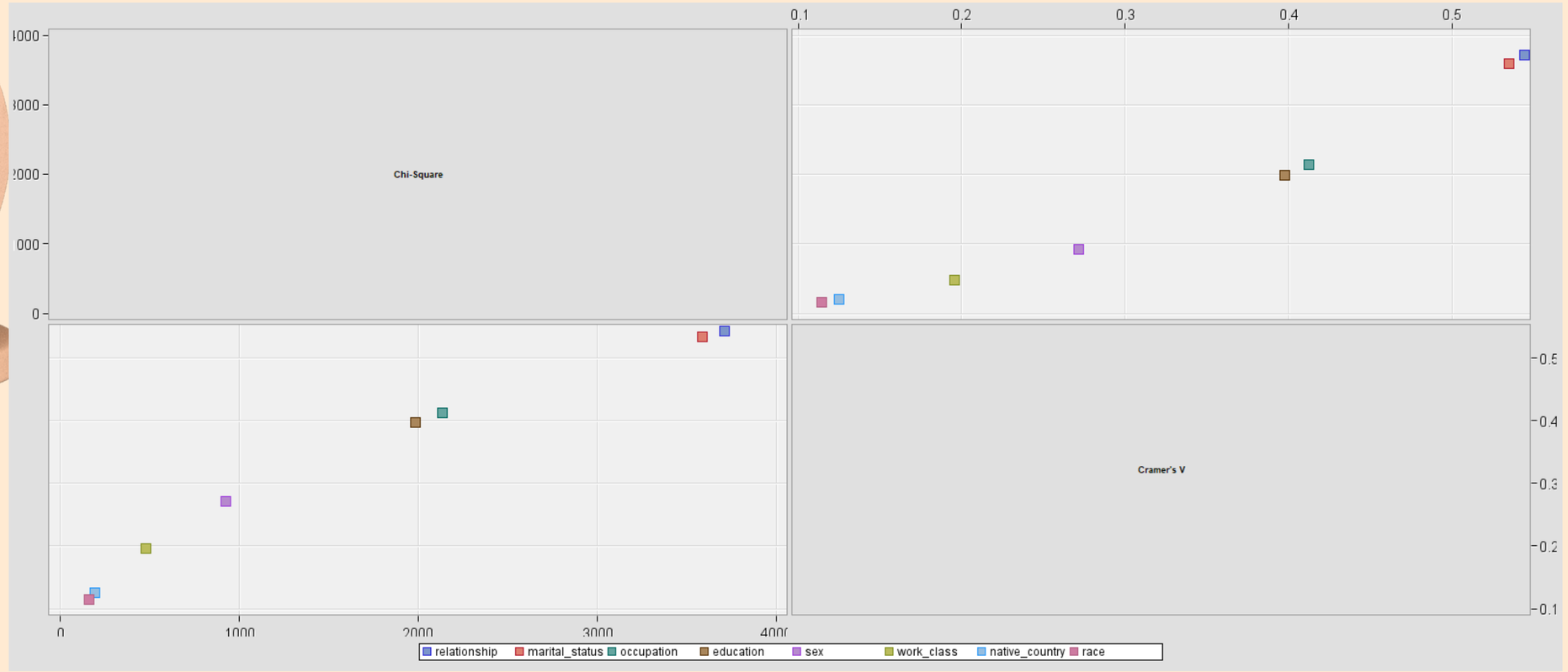
Output

```
40  *-------------------------------------------------------------*
41  * Report Output
42  *-------------------------------------------------------------*
43
44
45
46
47  Summary Statistics for Class Targets
48
49  Data=DATA
50
51                  Numeric    Formatted    Frequency
52    Variable      Value      Value        Count      Percent    Label
53
54  gross_income      .        <=50K         7841        50
55  gross_income      .        >50K          7841        50
56
57
58  Data=TRAIN
59
60                  Numeric    Formatted    Frequency
61    Variable      Value      Value        Count      Percent    Label
62
63  gross_income      .        <=50K         6272        50
64  gross_income      .        >50K          6272        50
65
66
67  Data=VALIDATE
68
69                  Numeric    Formatted    Frequency
70    Variable      Value      Value        Count      Percent    Label
71
72  gross_income      .        <=50K         1569        50
73  gross_income      .        >50K          1569        50
74
```

- Training set (80%):
  - 6272 instances from each target class
- Validation / Test set (20%):
  - 1569 instances from each target class

3.2 Explore

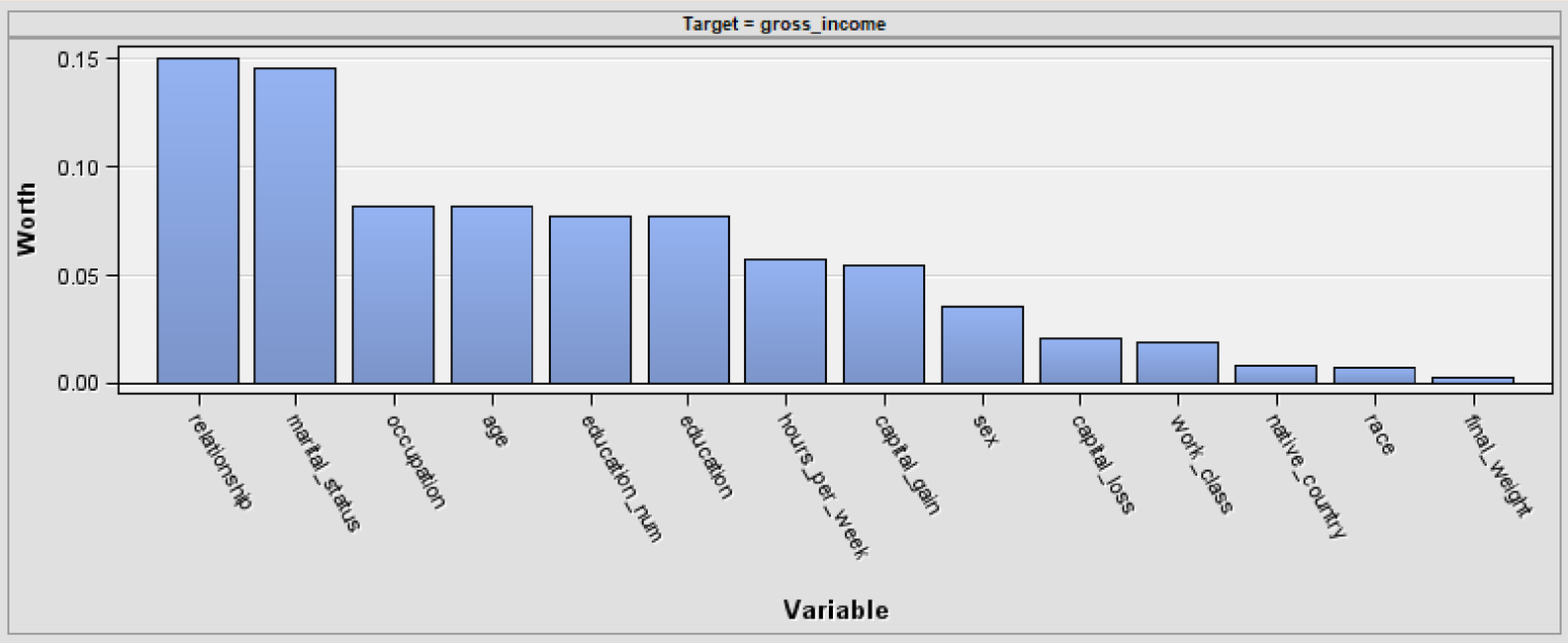
Chi-Square and Cramer's V matrix plot for all nominal variables.



relationship, marital_status, occupation, and education have strongest relationship with gross_income.

3.2 Explore

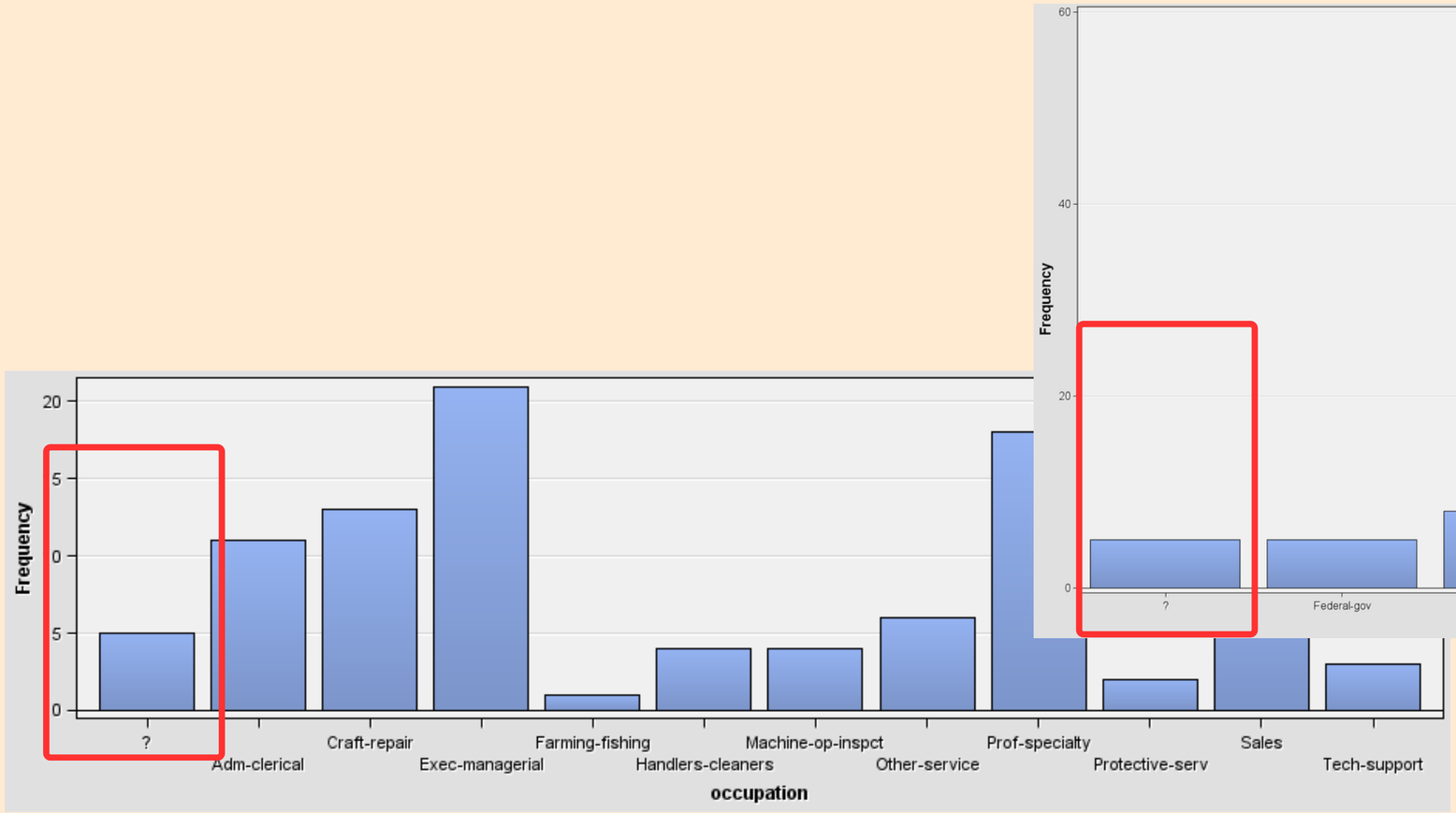
Worth of each independent variable for gross_income.

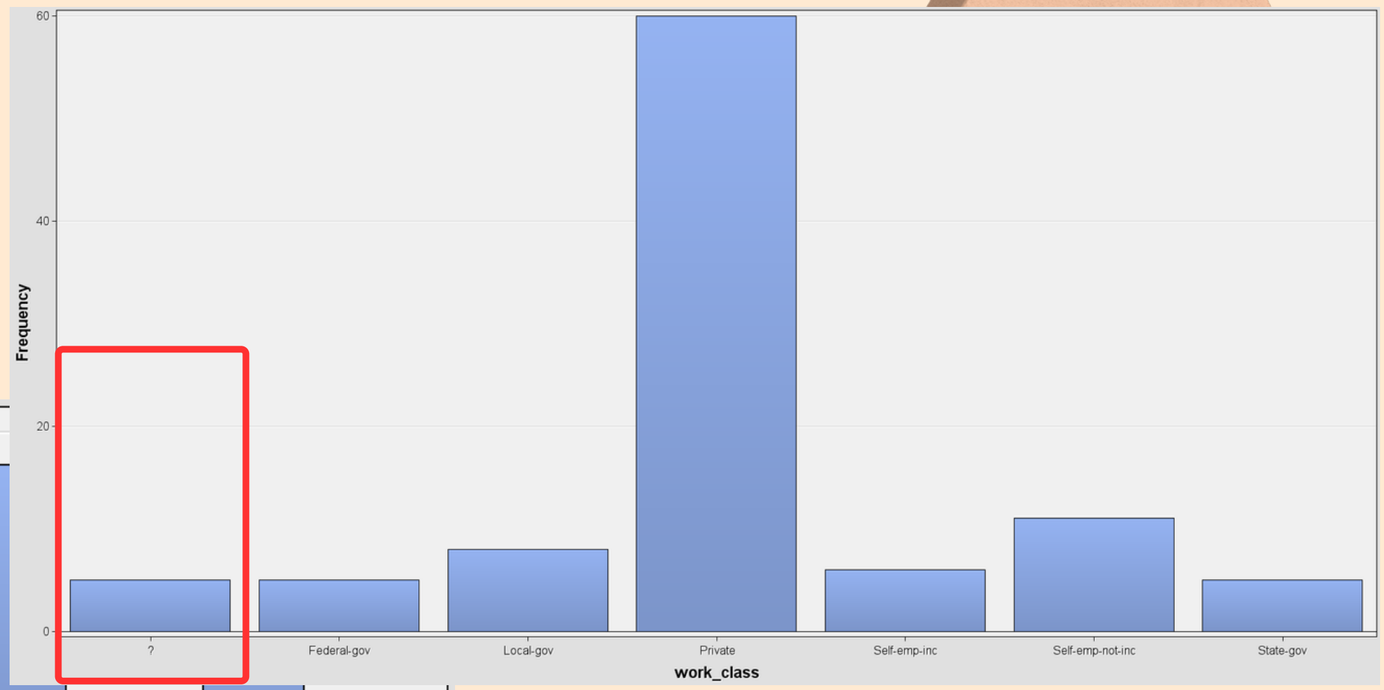


For interval variables, age, education_num, hours_per_week and capital_gain have the highest worth for predicting gross_income.

3.2 Explore

Bar plots of occupation and work_class columns.

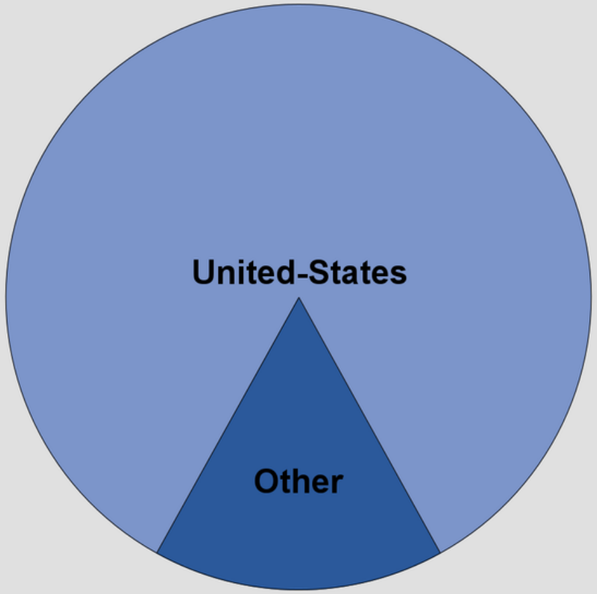




‘?’ in work_class and occupation columns will be assumed as missing values and will be inferenced using inference model.

3.2 Explore

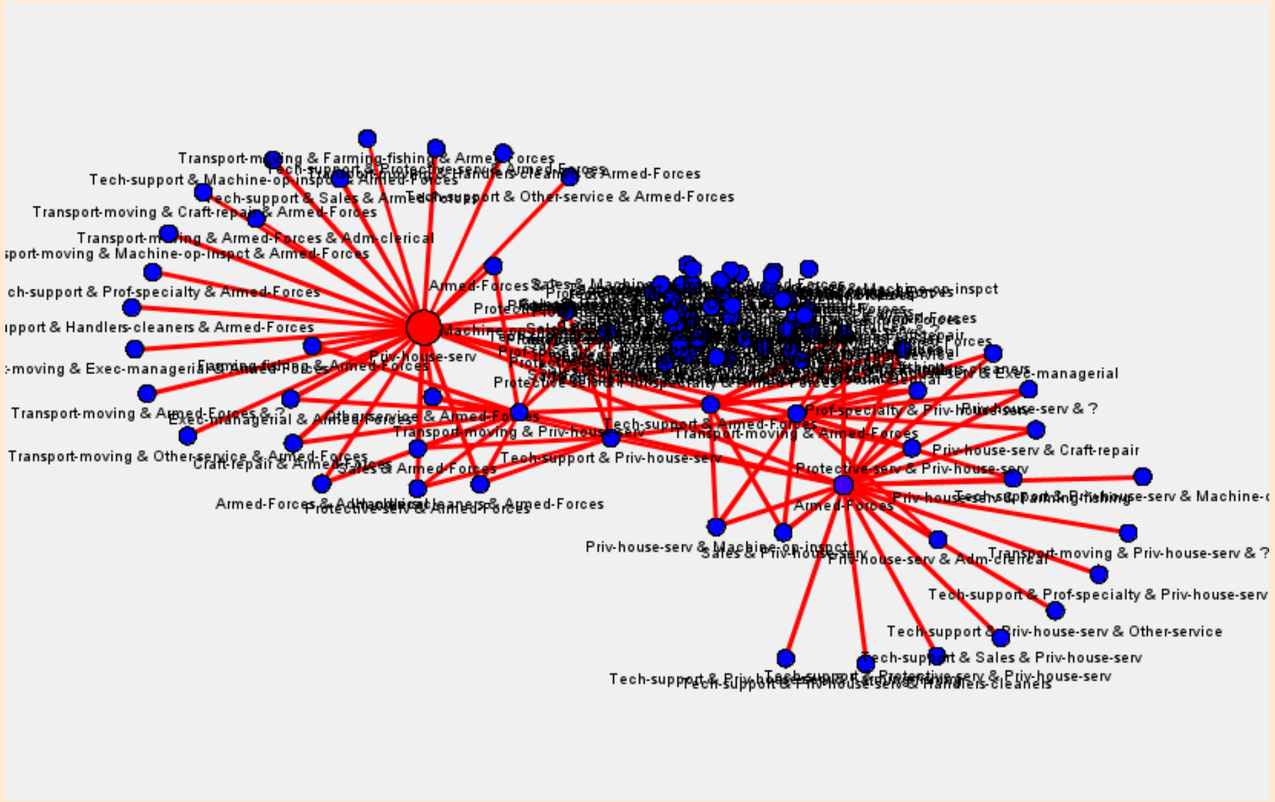
Pie chart showing distribution of native_country.



native_country columns will be divided into United-States and Other as countries other than ‘United-States’ only stand for a very small portion.

3.2 Explore

Association Rule Analysis



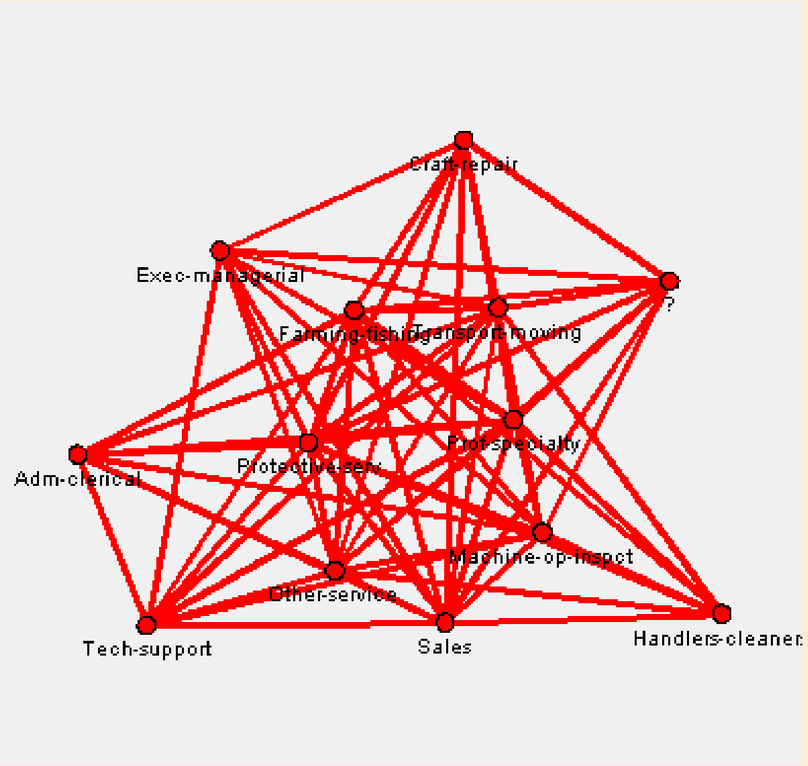
Link graph of association rules for occupation column.

Map	Rule
RULE1	Armed-Forces ==> Priv-house-serv
RULE2	Transport-moving & Armed-Forces ==> Priv-house-serv
RULE3	Tech-support & Armed-Forces ==> Priv-house-serv
RULE4	Sales & Armed-Forces ==> Priv-house-serv
RULE5	Protective-serv & Armed-Forces ==> Priv-house-serv
RULE6	Prof-specialty & Armed-Forces ==> Priv-house-serv
RULE7	Other-service & Armed-Forces ==> Priv-house-serv
RULE8	Machine-op-inspct & Armed-Forces ==> Priv-house-serv
RULE9	Handlers-cleaners & Armed-Forces ==> Priv-house-serv
RULE10	Farming-fishing & Armed-Forces ==> Priv-house-serv

Association rule table for occupation column.

3.2 Explore

Sequence Analysis



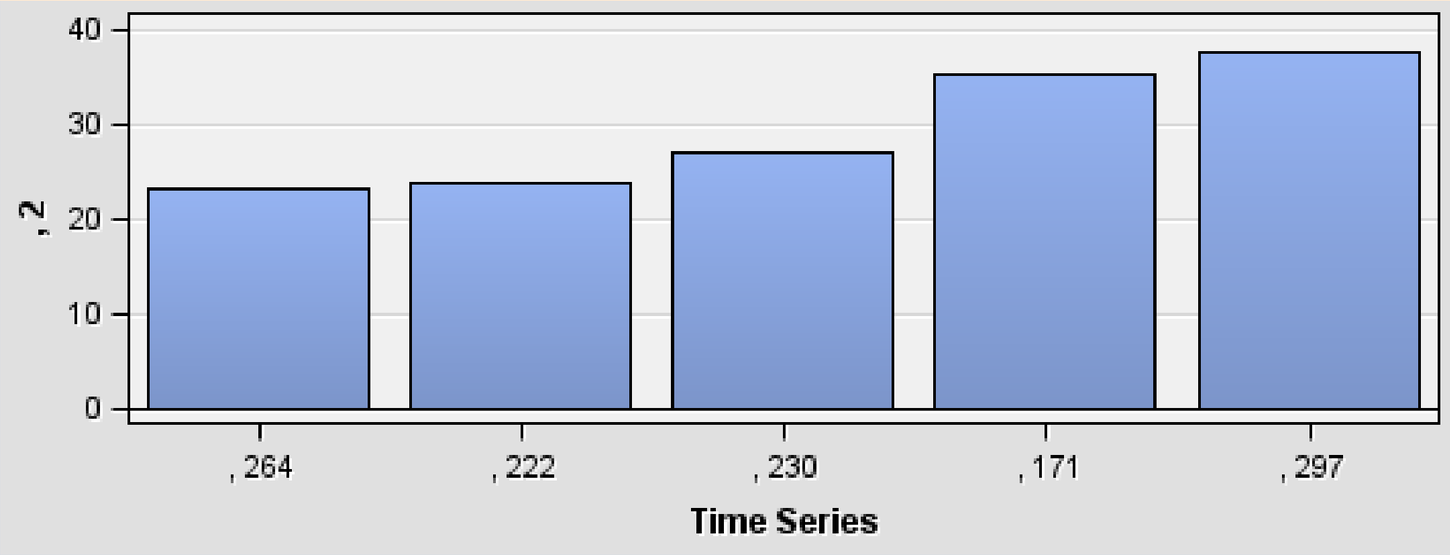
Link graph of sequence analysis for occupation column.

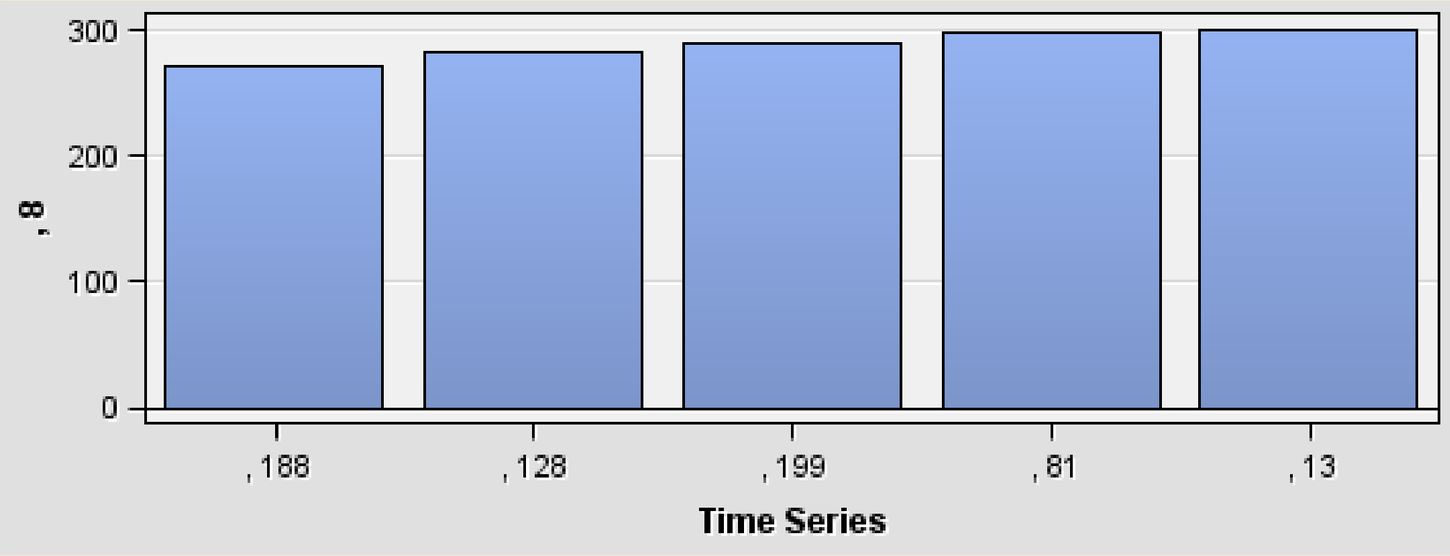
Map	Rule
RULE1	? ==> ?
RULE2	Adm-clerical ==> ?
RULE3	Craft-repair ==> ?
RULE4	Exec-managerial ==> ?
RULE5	Farming-fishing ==> ?
RULE6	Handlers-cleaners ==> ?
RULE7	Machine-op-inspct ==> ?
RULE8	Other-service ==> ?
RULE9	Prof-specialty ==> ?
RULE10	Protective-serv ==> ?
RULE11	Sales ==> ?
RULE12	Transport-moving ==> ?
RULE13	? ==> Adm-clerical
RULE14	Adm-clerical ==> Adm-clerical
RULE15	Craft-repair ==> Adm-clerical

Sequence analysis rule table for occupation column.

3.2 Explore

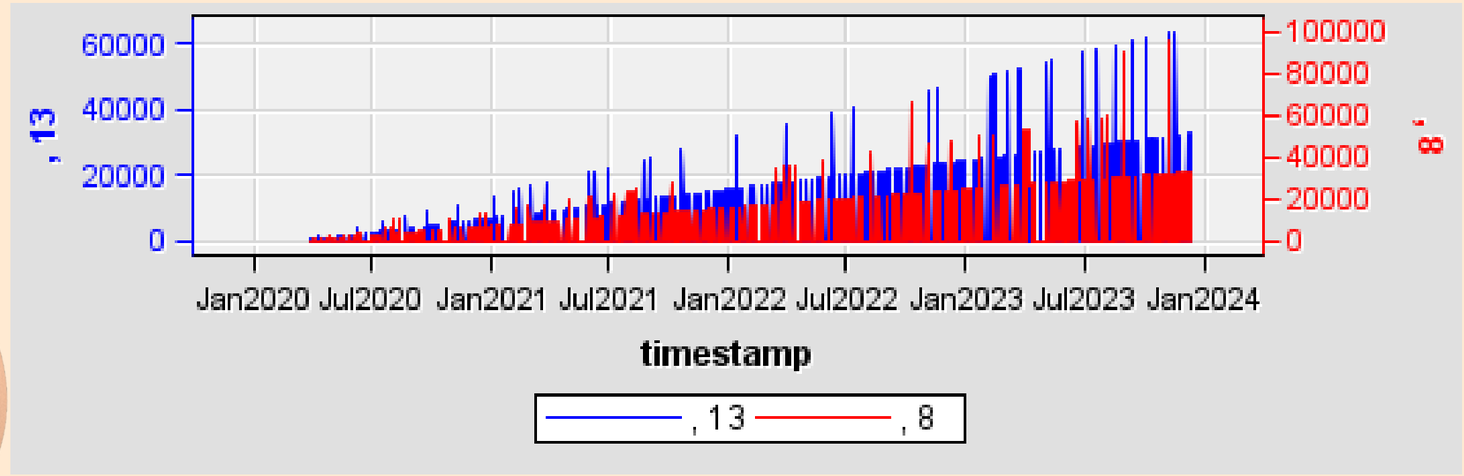
Time Series Clustering



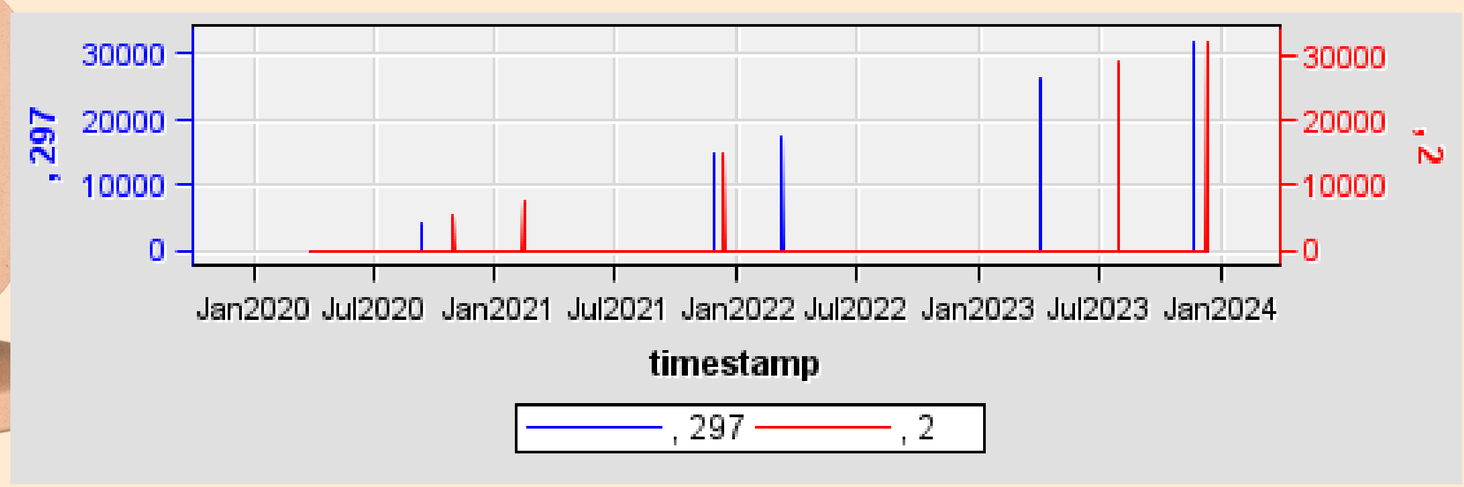


Time series clustering has been done on the marital_status, occupation and relationship for capital_gain to find out similar time series of capital gain.

3.2 Explore



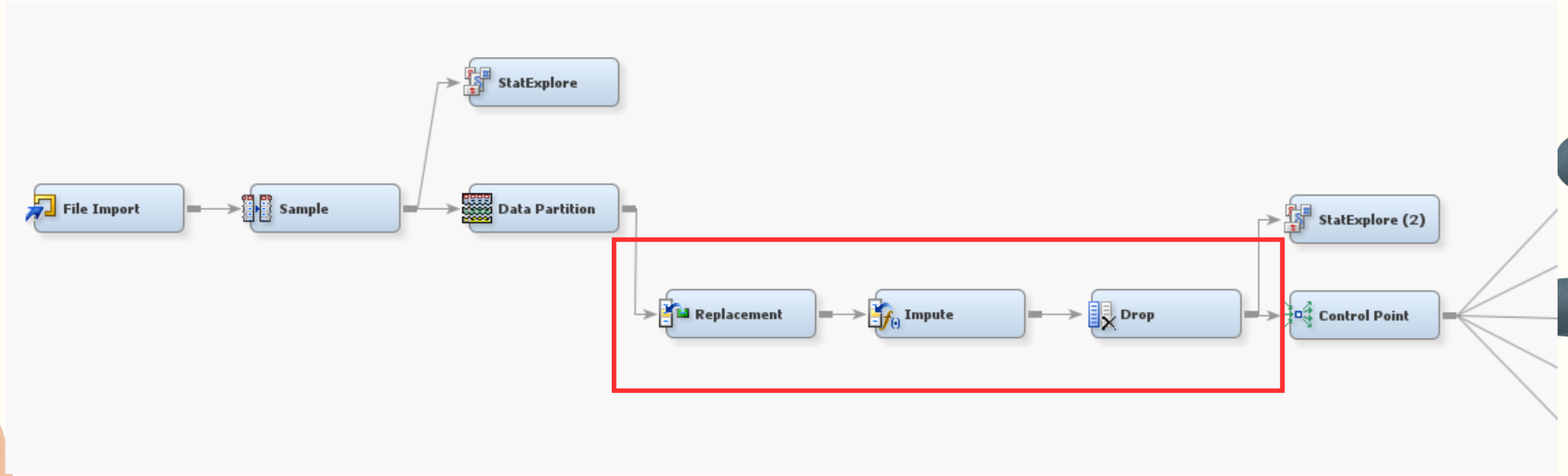
Plot of time series that is most similar to ts-8, the ts-13.



Plot of time series that is most similar to ts-2, the ts-297.

3.3 Modify

Overview



Our implementation “Modify” phase based on SEMMA methodology includes 3 steps:

- Replacement
- Imputation
- Deletion of unused variables

3.3 Modify

Overview

Variable	Actions to be taken
native_country	Replace countries other than “United-States” as “Others”, replace “?” with “_MISSING_”, impute “_MISSING_” with mode (“United-States”).
occupation	Replace “?” with “_MISSING_”, impute “_MISSING_” with inference algorithm.
work_class	Replace “?” with “_MISSING_”, impute “_MISSING_” with inference algorithm.
capital_gain	Drop variable, most values are “0” that don’t convey significant insight.
capital_loss	Drop variable, most values are “0” that don’t convey significant insight.
education	Drop variable, since “education_year” contribute the same meaning.
final_weight	Drop variable, low variable worth for predicting income group.
relationship	Drop variable, “marital-status” and “gender” are more representative and contribute the similar meaning as this variable.

Steps to be taken based on exploratory data analysis (EDA)

3.3 Modify

Replacement

native_country	?	_MISSING_
native_country	Mexico	Others
native_country	Philippines	Others
occupation	?	_MISSING_
occupation	Handlers-cleaners	
work_class	?	_MISSING_

Edit Variables:

- Replace '?' with '_MISSING_'
- Replace countries other than United-States as 'Others'

Result: missing values are replaced

Variable	Role	Train
native_country	INPUT	1232
occupation	INPUT	564
work_class	INPUT	563

	Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value
34	native_country	?	C	?	.	_blank_
35	native_country	Mexico	C	Mexico	.	Others
36	native_country	Philippines	C	Philippines	.	Others
37	native_country	Germany	C	Germany	.	Others
38	native_country	Canada	C	Canada	.	Others
39	native_country	India	C	India	.	Others

3.3 Modify

Imputation

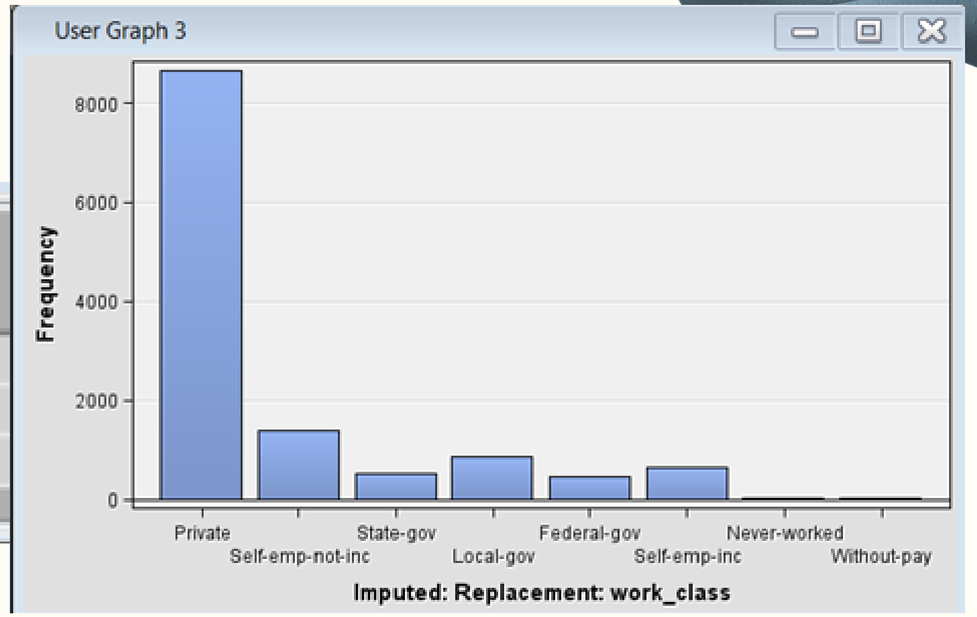
Edit Variables:

Name	Use	Method	Use Tree
REP_native_cou	Default	Default	Default
REP_occupation	Yes	Tree Surrogate	Yes
REP_work_class	Yes	Tree Surrogate	Yes

Change method to tree surrogate for REP_occupation and REP_work_class

Result:

Variable Name	Impute Method	Imputed Variable	Impute Value
REP_native_country	COUNT	IMP_REP_native_c...	United-States
REP_occupation	TREESURR	IMP_REP_occupati...	
REP_work_class	TREESURR	IMP_REP_work_cl...	



3.3 Modify

Deletion of Variables

Name	Drop	Role	Level
IMP_REP_native	**Default**	Input	Nominal
IMP_REP_occup	**Default**	Input	Nominal
IMP_REP_work_	**Default**	Input	Nominal
WARN	**Default**	Assessment	Nominal
dataobs	**Yes**	ID	Interval
age	**Default**	Input	Interval
capital_gain	**Yes**	Input	Interval
capital_loss	**Yes**	Input	Interval
education	**Yes**	Input	Nominal
education_num	**Default**	Input	Interval
final_weight	**Yes**	Input	Interval
gross_income	**Default**	Target	Binary
hours_per_week	**Default**	Input	Interval
marital_status	**Default**	Input	Nominal
native_country	**Yes**	Rejected	Nominal
occupation	**Yes**	Rejected	Nominal
race	**Default**	Input	Nominal
relationship	**Yes**	Input	Nominal
sex	**Default**	Input	Nominal
work_class	**Yes**	Rejected	Nominal

Name	Use	Report	Role	Level
IMP_REP_native	**Default**	**No**	Input	Nominal
IMP_REP_occup	**Default**	**No**	Input	Nominal
IMP_REP_work_	**Default**	**No**	Input	Nominal
age	**Default**	**No**	Input	Interval
education_num	**Default**	**No**	Input	Interval
gross_income	**Default**	**No**	Target	Binary
hours_per_week	**Default**	**No**	Input	Interval
marital_status	**Default**	**No**	Input	Nominal
race	**Default**	**No**	Input	Nominal
sex	**Default**	**No**	Input	Nominal

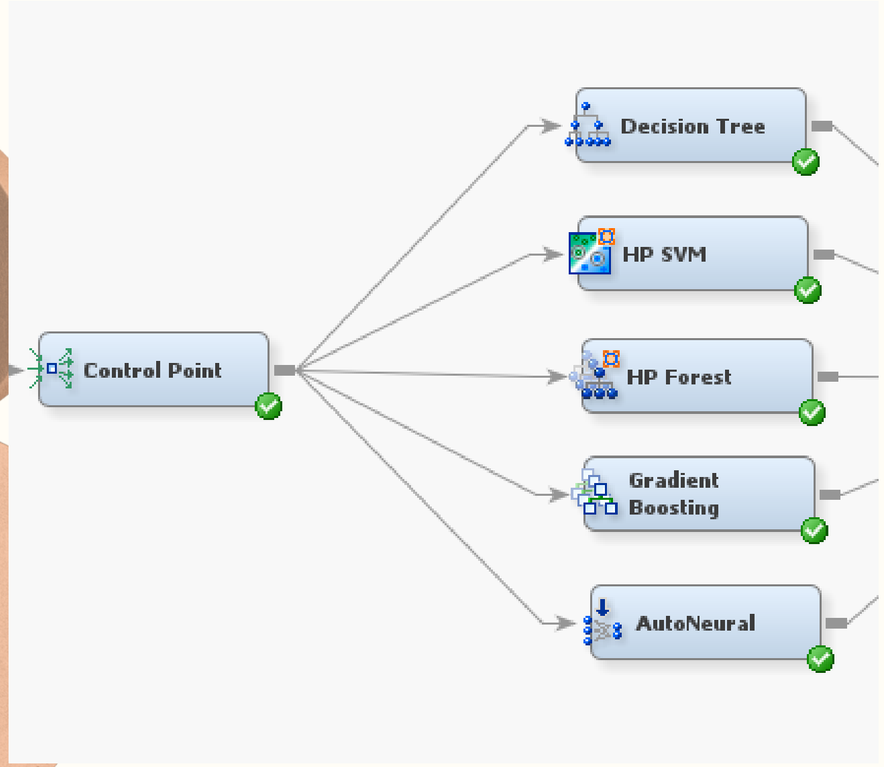
Drop Variables:

- Based on EDA:
  - capital_gain, capital_loss, education, final_weight, relationship

- By-product of “imputation” and “replacement”:
  - _dataobs_, native_country, occupation, work_class

3.4 Model

Decision Tree



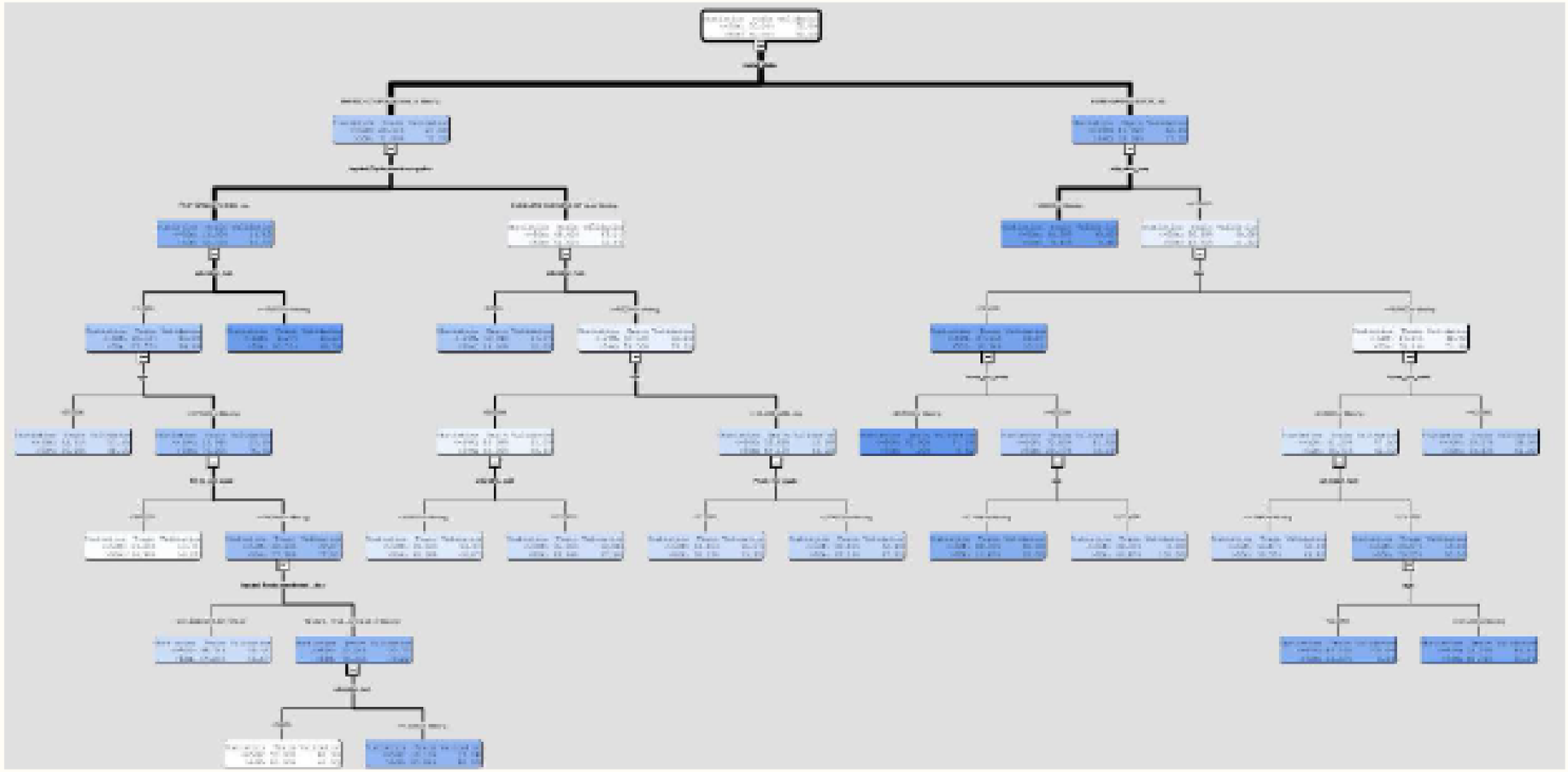
Create nodes and connect to “Control Point” node

Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	

Set maximum depth as '10' as we have 9 input variables

3.4 Model

Decision Tree



Overview of Decision Tree created

3.4 Model

Support Vector Machine (SVM)

General	
Node ID	HPSVM
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Maximum Iterations	25
Use Missing as Level	No
Tolerance	1.0E-6
Penalty	1.0
Optimization Method	
Optimization Method	Interior Point
Interior Point Options	
Active Set Options	
Status	
Create Time	1/20/24 10:25 AM
Run ID	2010b0f7-6d34-4b45-9d3f-1b2987b7545f
Last Error	
Last Status	Complete
Last Run Time	1/20/24 10:29 AM
Run Duration	0 Hr. 0 Min. 8.53 Sec.
Grid Host	
User-Added Node	No

Use default parameter value as
SAS Enterprise Miner

Random Forest

Train	
Variables	
Tree Options	
Maximum Number of Trees	40
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split Sea	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000

Set maximum number of trees
to 40 to prevent overfitting

3.4 Model

Gradient Boosting

Train	
Variables	...
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Misclassification
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	

Set “Assessment Measure” to Misclassification

AutoNeural

Train	
Variables	...
Model Options	
Architecture	Single Layer
Termination	Overfitting
Train Action	Search
Target Layer Error Function	Default
Maximum Iterations	8
Number of Hidden Units	2
Tolerance	Medium
Total Time	One Hour
Increment and Search Options	
Adjust Iterations	Yes
Freeze Connections	No
Total Number of Hidden Units	30
Final Training	Yes
Final Iterations	5
Activation Functions	
Direct	Yes
Exponential	No
Identity	No
Logistic	No
Normal	Yes
Reciprocal	No
Sine	Yes
Softmax	No
Square	No
Tanh	Yes
Score	
Hidden Units	No
Residuals	Yes
Standardization	No

Use default parameter value as SAS Enterprise Miner

3.5 Assess

Misclassification Rate

Selected Model	Model Node	Model Description	Valid: Misclassification Rate
Y	Boost	Gradient Boosting	0.19216
	HPDMForest	HP Forest	0.19503
	AutoNeural	AutoNeural	0.19981
	Tree	Decision Tree	0.20714
	HPSVM	HP SVM	0.21256

Best performing model is Gradient Boosting with a misclassification rate of 0.192

3.5 Assess

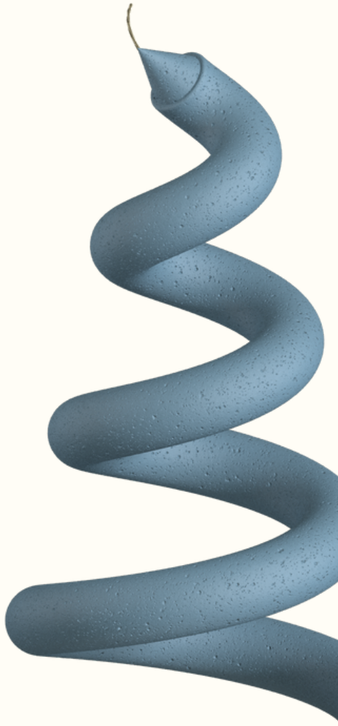
Classification metric

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	TRAIN	gross_income		1177	5082	1190	5095
Tree	Decision Tree	VALIDATE	gross_income		328	1247	322	1241
HPSVM	HP SVM	TRAIN	gross_income		1054	4644	1628	5218
HPSVM	HP SVM	VALIDATE	gross_income		267	1169	400	1302
HPDMForest	HP Forest	TRAIN	gross_income		942	4928	1344	5330
HPDMForest	HP Forest	VALIDATE	gross_income		257	1214	355	1312
Boost	Gradient Boosting	TRAIN	gross_income		912	4821	1451	5360
Boost	Gradient Boosting	VALIDATE	gross_income		230	1196	373	1339
AutoNeural	AutoNeural	TRAIN	gross_income		811	4524	1748	5461
AutoNeural	AutoNeural	VALIDATE	gross_income		202	1144	425	1367

Classification metric for all models

3.5 Assess

F1-Score

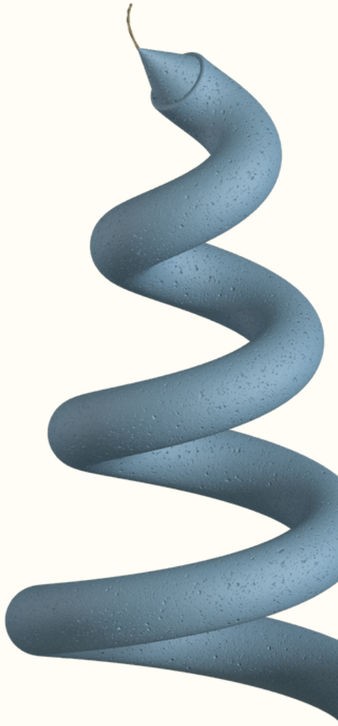


Model	F1-score
Decision Tree	0.792
HP SVM	0.796
HP Forest	0.811
Gradient Boosting	0.816
AutoNeural	0.814

The best F1-score is Gradient Boosting, with F1-score of 0.816

3.5 Assess

Precision



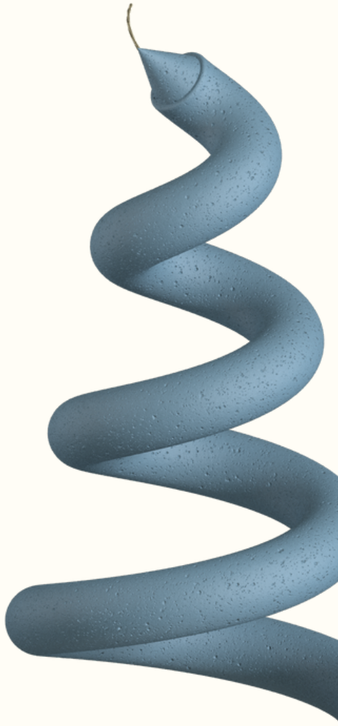
Model	Precision
Decision Tree	0.794
HP SVM	0.765
HP Forest	0.787
Gradient Boosting	0.782
AutoNeural	0.763

Decision tree had the highest precision of 0.794

3.5 Assess

Recall

Model	Recall
Decision Tree	0.791
HP SVM	0.830
HP Forest	0.836
Gradient Boosting	0.853
AutoNeural	0.871



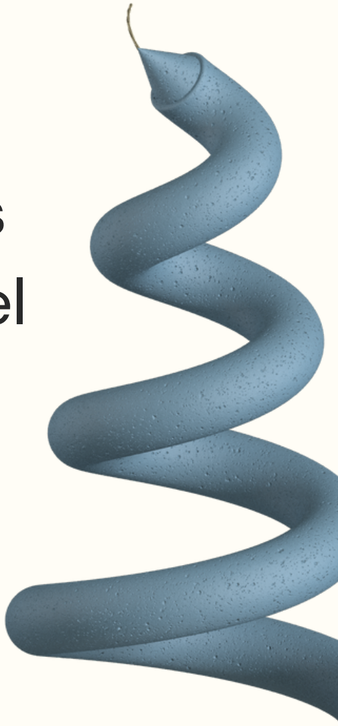
AutoNeural had the highest recall of 0.871

Conclusion

- Preprocessing: adding column names, handling missing values
- Trained 5 machine learning algorithms, best performing model is Gradient boosting

Future Works

- Dataset can include more balanced data
- More machine learning models can be fit to compare more models
- Can have hyperparameter tuning



Thank you