USER MANUAL GUIDE FOR CREATING A LOGISTIC MODEL

DR BRAIN STATS

HOME

Welcome to Dr Brain Stats

Your place for the supporting material for the books

Please click the links below to download the ZIP files, or the books to direct you to Amazon

Applied Predictive Modelling for Direct Marketing in SAS

Applied Predictive Modelling for Direct Marketing in R

Applied Predictive Modelling for Direct Marketing in Python

Introduction to applied statistics using Excel and SAS

Introduction to applied statistics using Excel and R

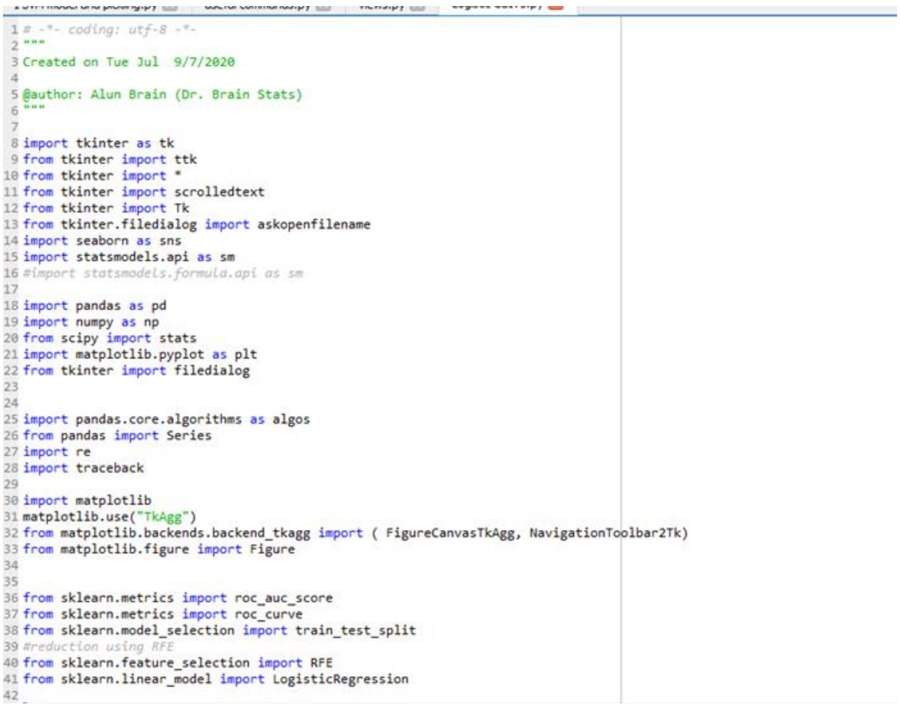
Statistical Analysis in Excel and Python

Learning SQL in Python/R

www.drbrainstats.com

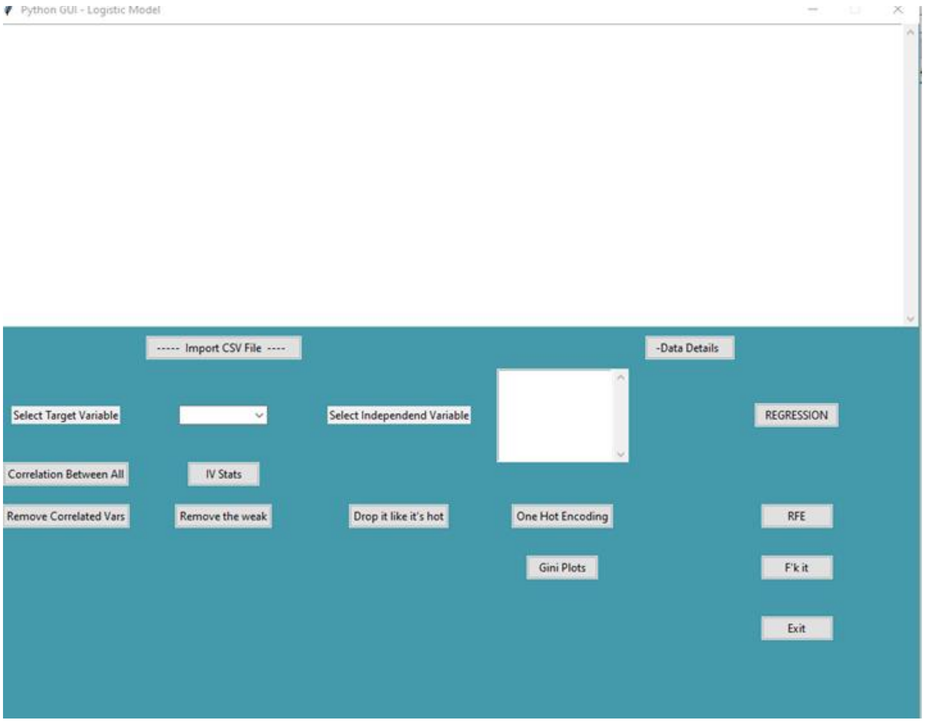
Instructions to running GLM GUI

1) First run all of the libraries



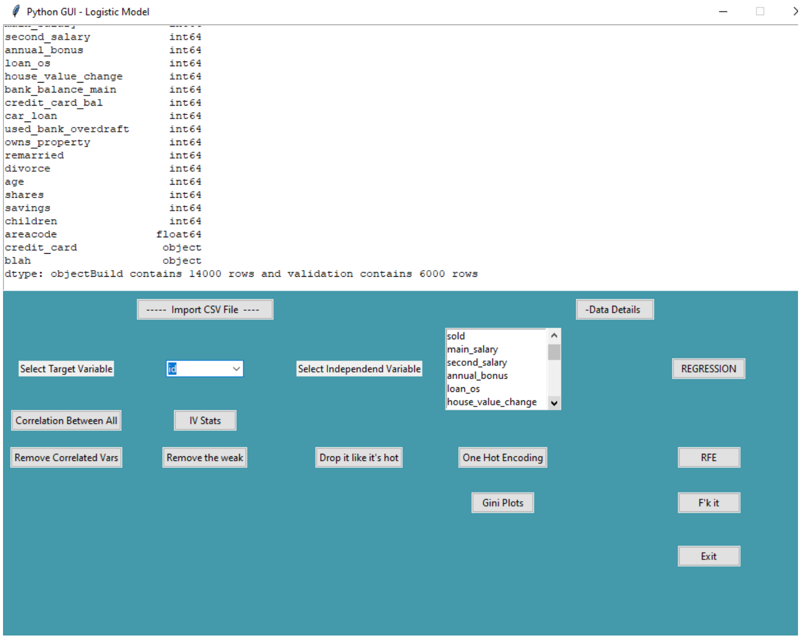
```
# -*- coding: utf-8 -*-
"""
Created on Tue Jul  9/7/2020

@author: Alun Brain (Dr. Brain Stats)
"""

import tkinter as tk
from tkinter import ttk
from tkinter import *
from tkinter import scrolledtext
from tkinter import Tk
from tkinter.filedialog import askopenfilename
import seaborn as sns
import statsmodels.api as sm
#import statsmodels.formula.api as sm

import pandas as pd
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
from tkinter import filedialog


import pandas.core.algorithms as algos
from pandas import Series
import re
import traceback

import matplotlib
matplotlib.use("TkAgg")
from matplotlib.backends.backend_tkagg import ( FigureCanvasTkAgg, NavigationToolbar2Tk)
from matplotlib.figure import Figure


from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
from sklearn.model_selection import train_test_split
#reduction using RFE
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression

```

2) Run all the code between:
    a. Tk().withdraw()
    b. root.mainloop()
3) You get the GUI



Using the GUI – Logistic regression

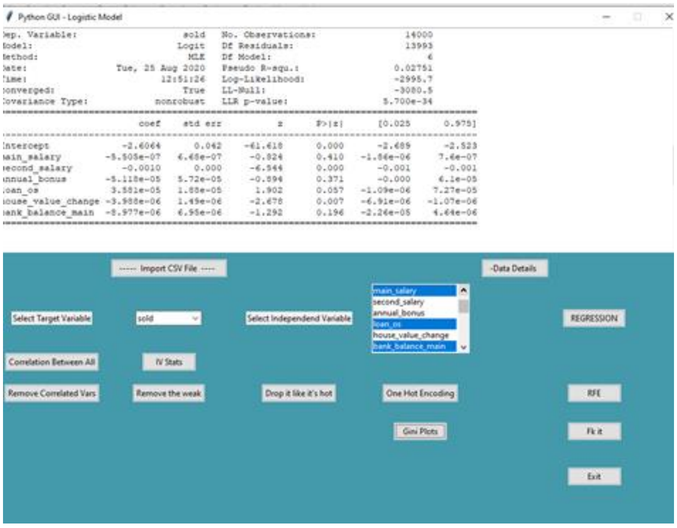
1) Import CSV file – click this to get your CSV file



```
second_salary             int64
annual_bonus              int64
loan_os                   int64
house_value_change        int64
bank_balance_main         int64
credit_card_bal           int64
car_loan                  int64
used_bank_overdraft       int64
owns_property             int64
remarried                 int64
divorce                   int64
age                       int64
shares                    int64
savings                   int64
children                  int64
areacode                float64
credit_card              object
blah                     object
dtype: objectBuild contains 14000 rows and validation contains 6000 rows
```

This will tell you information about your data and split you data in70/30 split, ashown

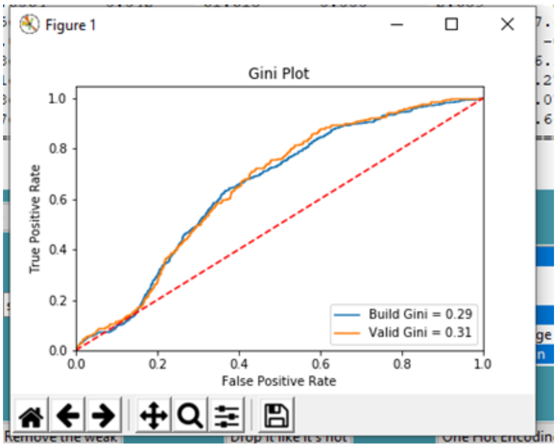
2) Create Initial Multi-linear regression
   - Select Target Variable
     - This can easily be changed, we will be using **sold**
   - Select Independent Variable
     - Click on each variable you wish to model, in this scenario , main_salary, loan_os andbank_balance_main
   - Then press **REGRESSION**



```
ep. Variable:                 sold   No. Observations:                14000
odel:                        Logit   Df Residuals:                    13993
ethod:                         MLE   Df Model:                            6
ate:              Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                      12:51:26   Log-Likelihood:                -2995.7
onverged:                     True   LL-Null:                       -3080.5
ovariance Type:          nonrobust   LLR p-value:                 5.700e-34
==========================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
Intercept           -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary       -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary        -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus      -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os            3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06  1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main -8.977e-06   6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
==========================================================================
```

Your model is in the screen.

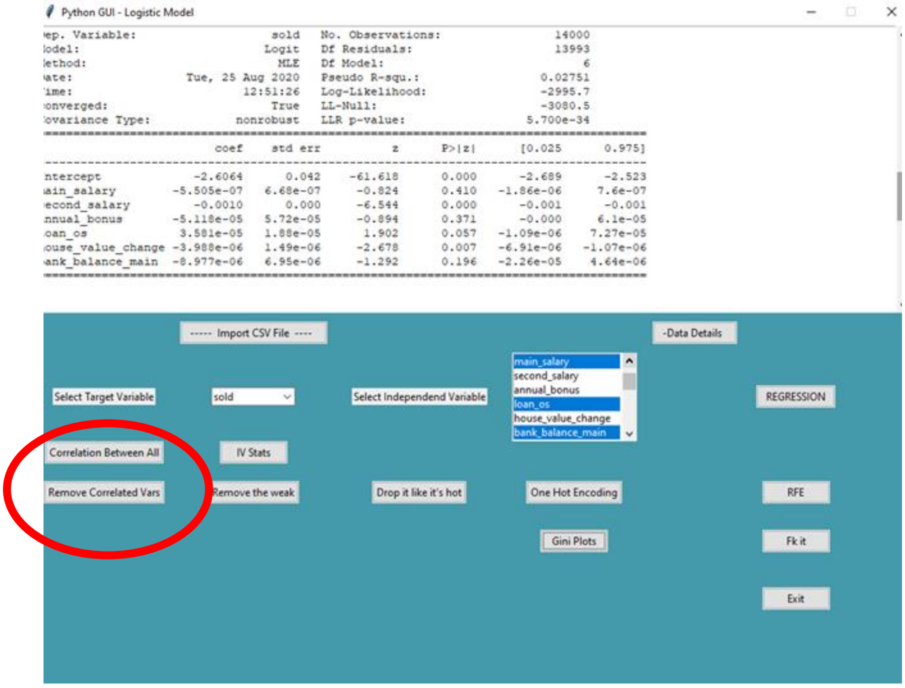
Clicking on Gini-plot gives you



It applies the model on to your holdout, so a buildb holdout comparison can be made.

Extra Funcationality

Correlation

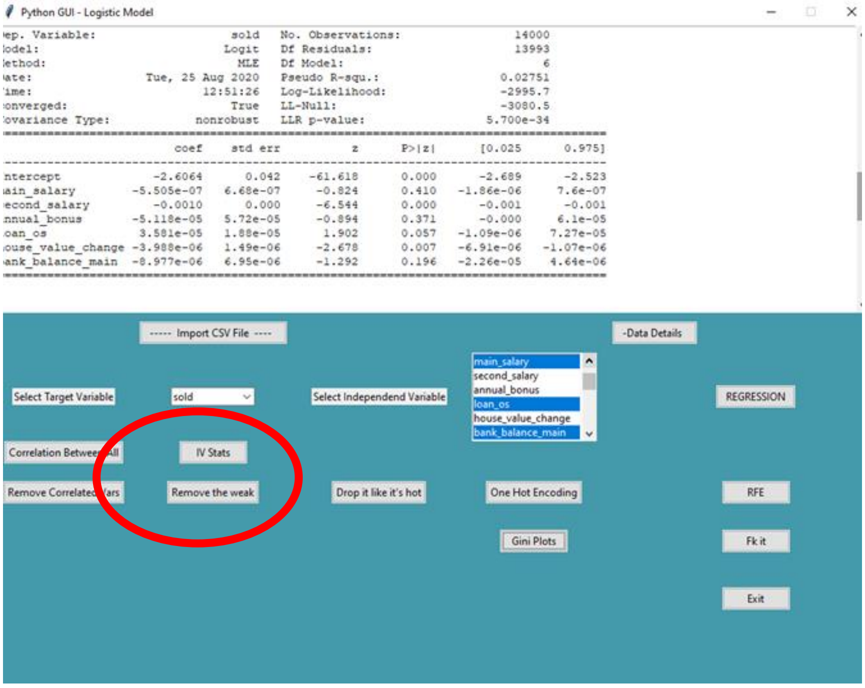


```
ep. Variable:                 sold   No. Observations:                14000
odel:                        Logit   Df Residuals:                    13993
ethod:                         MLE   Df Model:                            6
ate:              Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                      12:51:26   Log-Likelihood:                -2995.7
onverged:                     True   LL-Null:                       -3080.5
ovariance Type:          nonrobust   LLR p-value:                 5.700e-34
=========================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
ntercept             -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary        -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary         -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus       -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os             3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06   1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main  -8.977e-06   6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
=========================================================================================
```

Will provide all of your correlated statistics.

Will also remove any correlated values based on 0.7 cut-off

IV stats

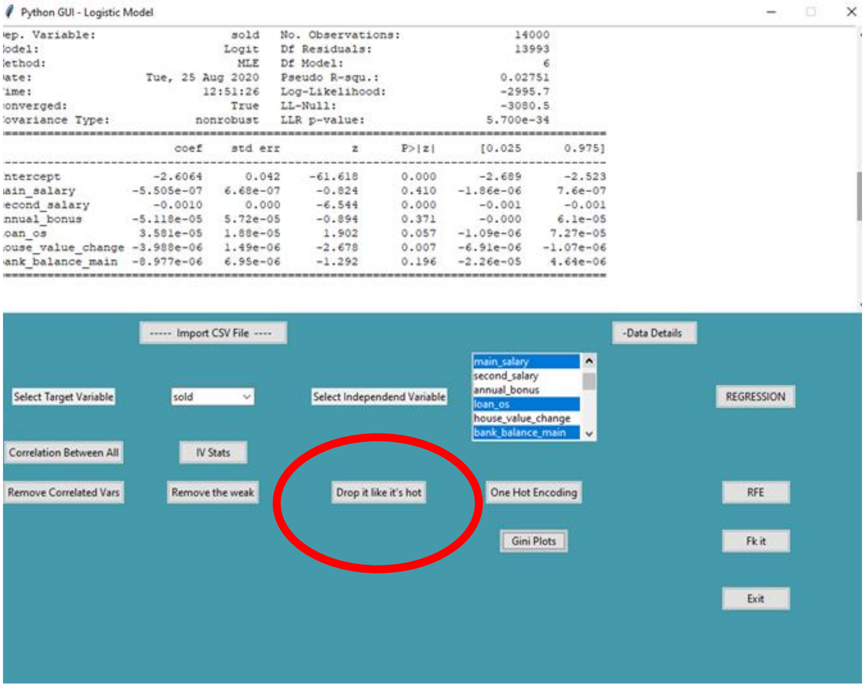


```
ep. Variable:                  sold   No. Observations:                14000
odel:                         Logit   Df Residuals:                    13993
ethod:                          MLE   Df Model:                            6
ate:               Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                       12:51:26   Log-Likelihood:                -2995.7
onverged:                      True   LL-Null:                       -3080.5
ovariance Type:           nonrobust   LLR p-value:                 5.700e-34
==============================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ntercept            -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary       -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary        -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus      -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os            3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06   1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main -8.977e-06   6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
==============================================================================
```

Will provide all of your IV statistics.

Will also remove any variables with an IV lower than 0.1

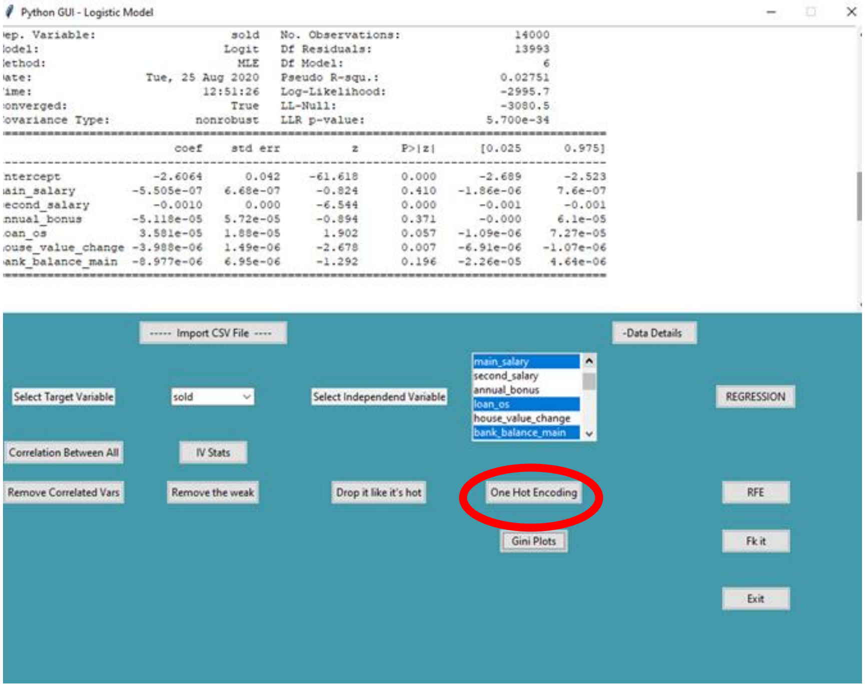
Drop it like its hot



```
ep. Variable:                  sold   No. Observations:                14000
odel:                         Logit   Df Residuals:                    13993
ethod:                          MLE   Df Model:                            6
ate:               Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                       12:51:26   Log-Likelihood:                -2995.7
onverged:                      True   LL-Null:                       -3080.5
ovariance Type:           nonrobust   LLR p-value:                 5.700e-34
==============================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ntercept            -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary       -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary        -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus      -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os            3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06   1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main -8.977e-06   6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
==============================================================================
```

Removes those categorical variables with over 10 bins (easily amendable). This is recommended before you use **One Hot encoding**

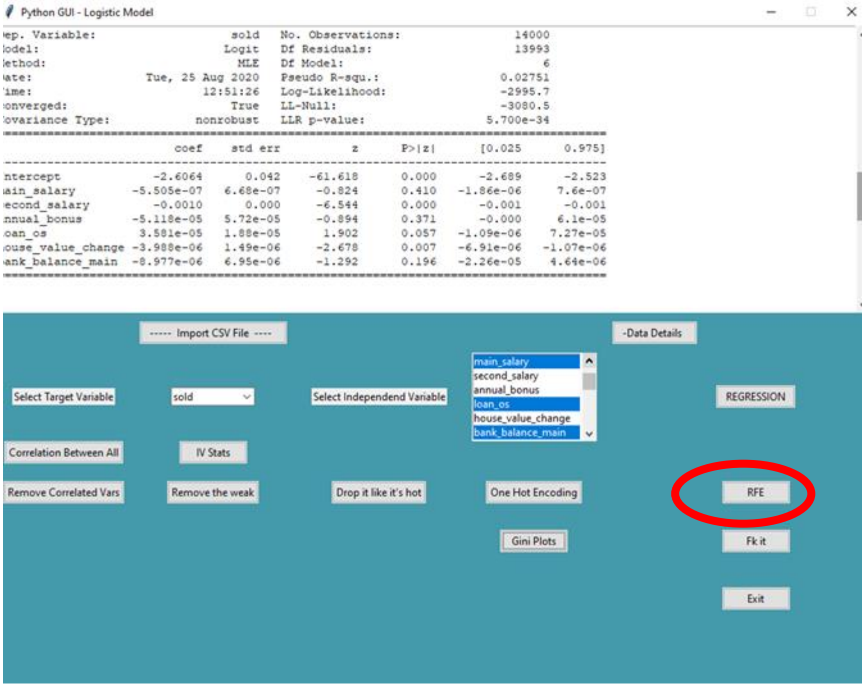
One Hot Encoding



```
ep. Variable:                   sold   No. Observations:                14000
odel:                          Logit   Df Residuals:                    13993
ethod:                           MLE   Df Model:                            6
ate:                Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                        12:51:26   Log-Likelihood:                -2995.7
onverged:                       True   LL-Null:                       -3080.5
ovariance Type:            nonrobust   LLR p-value:                 5.700e-34
=====================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
ntercept              -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary         -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary          -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus        -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os              3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06   1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main  -8.977e-06   6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
=====================================================================================
```

Create dummy variables on categorical variables, but removes one bin from each variable so it can not suffer from collinearity.

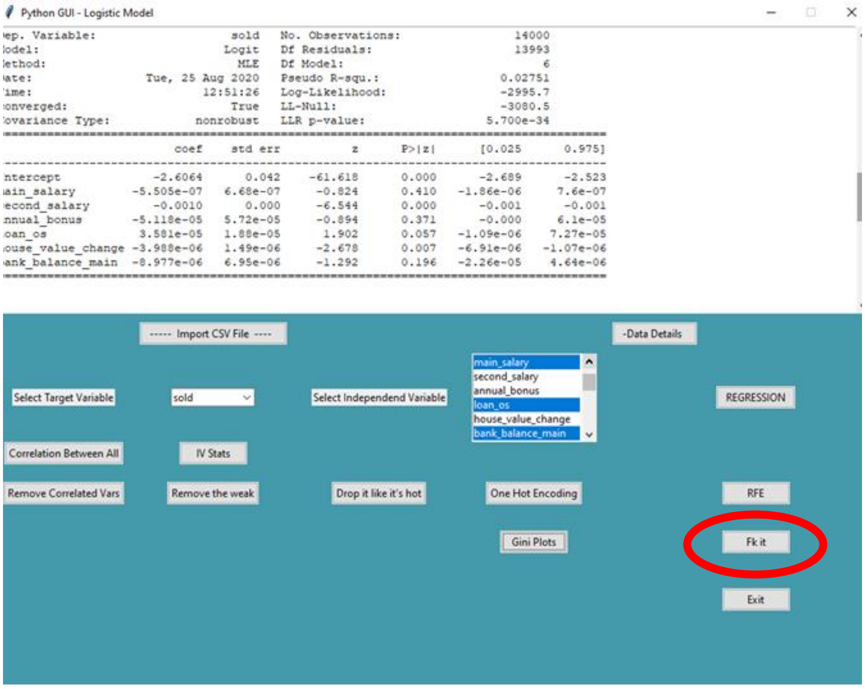
RFE



```
ep. Variable:                   sold   No. Observations:                14000
odel:                          Logit   Df Residuals:                    13993
ethod:                           MLE   Df Model:                            6
ate:                Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                        12:51:26   Log-Likelihood:                -2995.7
onverged:                       True   LL-Null:                       -3080.5
ovariance Type:            nonrobust   LLR p-value:                 5.700e-34
=====================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
ntercept              -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary         -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary          -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus        -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os              3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06   1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main  -8.977e-06   6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
=====================================================================================
```

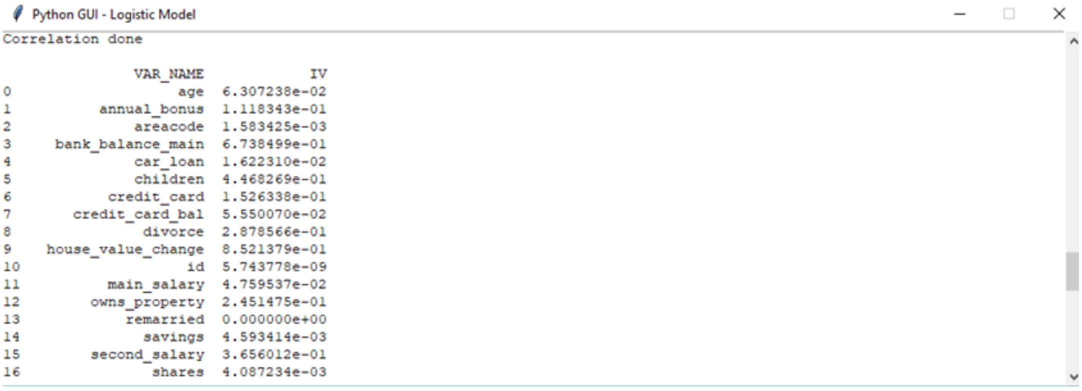
Provides you with your top performing variables, but now variable reduction is done as this part is subjective.

JUST GIVE ME A MODEL BUTTON



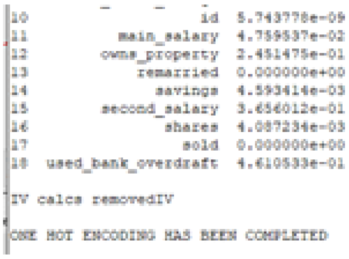
```
ep. Variable:                   sold   No. Observations:                14000
odel:                          Logit   Df Residuals:                    13993
ethod:                           MLE   Df Model:                            6
ate:                Tue, 25 Aug 2020   Pseudo R-squ.:                 0.02751
ime:                        12:51:26   Log-Likelihood:                -2995.7
onverged:                       True   LL-Null:                       -3080.5
ovariance Type:            nonrobust   LLR p-value:                 5.700e-34
==================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
ntercept           -2.6064      0.042    -61.618      0.000      -2.689      -2.523
ain_salary      -5.505e-07   6.68e-07     -0.824      0.410   -1.86e-06     7.6e-07
econd_salary       -0.0010      0.000     -6.544      0.000      -0.001      -0.001
nnual_bonus     -5.118e-05   5.72e-05     -0.894      0.371      -0.000     6.1e-05
oan_os           3.581e-05   1.88e-05      1.902      0.057   -1.09e-06    7.27e-05
ouse_value_change -3.988e-06 1.49e-06     -2.678      0.007   -6.91e-06   -1.07e-06
ank_balance_main -8.977e-06  6.95e-06     -1.292      0.196   -2.26e-05    4.64e-06
==================================================================================
```

This conducts variable reduction, one hot encoding, model build and graph in one go



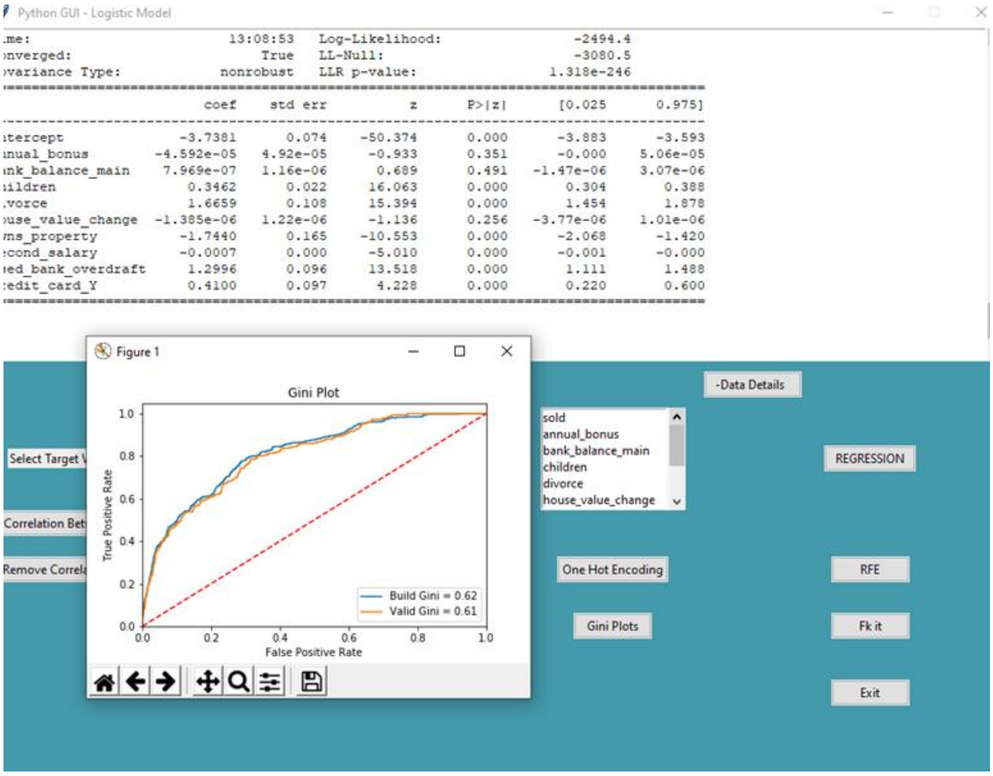
```
Correlation done

              VAR_NAME            IV
0                  age  6.307238e-02
1         annual_bonus  1.118343e-01
2             areacode  1.583425e-03
3    bank_balance_main  6.738499e-01
4             car_loan  1.622310e-02
5             children  4.468269e-01
6          credit_card  1.526338e-01
7      credit_card_bal  5.550070e-02
8              divorce  2.878566e-01
9    house_value_change  8.521379e-01
10                  id  5.743778e-09
11         main_salary  4.759537e-02
12       owns_property  2.451475e-01
13          remarried  0.000000e+00
14             savings  4.593414e-03
15       second_salary  3.656012e-01
16              shares  4.087234e-03
```

And so on…



```
10                  id  5.743778e-09
11         main_salary  4.759537e-02
12       owns_property  2.451475e-01
13           remarried  0.000000e+00
14             savings  4.593414e-03
15       second_salary  3.656012e-01
16              shares  4.087234e-03
17                sold  0.000000e+00
18  used_bank_overdraft  4.610533e-01

IV calcs removedIV

ONE HOT ENCODING HAS BEEN COMPLETED
```

And on….



```
.me:                    13:08:53   Log-Likelihood:             -2494.4
>nverged:                   True   LL-Null:                    -3080.5
>variance Type:        nonrobust   LLR p-value:              1.318e-246
==============================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
.tercept             -3.7381      0.074    -50.374      0.000      -3.883      -3.593
.nual_bonus       -4.592e-05   4.92e-05     -0.933      0.351      -0.000    5.06e-05
.nk_balance_main   7.969e-07   1.16e-06      0.689      0.491   -1.47e-06    3.07e-06
.ildren               0.3462      0.022     16.063      0.000       0.304       0.388
.vorce                1.6659      0.108     15.394      0.000       1.454       1.878
>use_value_change -1.385e-06   1.22e-06     -1.136      0.256   -3.77e-06    1.01e-06
>ns_property         -1.7440      0.165    -10.553      0.000      -2.068      -1.420
.cond_salary         -0.0007      0.000     -5.010      0.000      -0.001      -0.000
>ed_bank_overdraft    1.2996      0.096     13.518      0.000       1.111       1.488
.edit_card_Y          0.4100      0.097      4.228      0.000       0.220       0.600
==============================================================================
```