```
/*This coding is to prepare the final dataset for fraud detection models that
use either 28 raw accounting variables or 14 financial ratios as features in
the prediction model;*/
/*All the accounting raw data are downloaded from WRDS/COMPUSTAT/2017.4.3--
fraud.compustatindustrial7815;*/
/*AAER_firm_year dataset has two sources (see our shared dataset): in
addition to AAER 20160930 database we purchased from the University of
California-Berkeley’s Center for Financial Reporting and Management (CFRM),
we hand collect AAER frauds from SEC website in the period from 20160930 to
20181231*/
/*Note: we use CIK as company ID first to merge the data with COMPUSTAT
because the CFRM AAER database 20160930 only provides CIK as firm ID;*/

PROC SQL;
CREATE TABLE temp1 AS SELECT
a.*, b.yeara, b.p_aaer,b.understatement
FROM fraud.compustatindustrial7815 AS a LEFT JOIN fraud.aaer_firm_year AS b
ON a.cik=b.cik
AND a.fyear=b.yeara;
QUIT;
RUN;

PROC SORT DATA=temp1 NODUPKEY;
BY gvkey datadate;
RUN;

DATA temp1; SET temp1;
IF yeara=. THEN misstate=0;
ELSE IF yeara^=. THEN misstate=1;
RUN;

DATA temp1; SET temp1;
IF understatement=. THEN understatement=0;
RUN;

data temp1; set temp1;
if 1978<=fyear<=2015;
run;

DATA temp1; SET temp1;
IF 6000<=sich<=6999 THEN INSBNK=1;
ELSE INSBNK=0 ;
RUN;

proc sort data=temp1 nodupkey;
by gvkey fyear;
run;

DATA temp1; SET temp1;
IF MISSING (at) THEN DELETE;
RUN;
```

```
/*calculate 14 financial ratios;*/
DATA temp1; SET temp1;
lag_gvkey=LAG(gvkey);
lag_fyear=LAG(fyear);
lag_cik=LAG(cik);
lag_at=LAG(at);
RUN;

DATA temp1; SET temp1;
IF lag_gvkey^=gvkey THEN DO;
lag_fyear=.;
lag_cik=.;
lag_at=.;
END;
RUN;

*changes in working capital accruals;
DATA temp1; SET temp1;
IF MISSING(txp) THEN txp=0;
wc=(act-che)-(lct-dlc-txp);
lag_wc=LAG(wc);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN ch_wc=wc-lag_wc;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_wc=.;
RUN;
DATA temp1; SET temp1;
dch_wc=ch_wc*2/(at+lag_at);
RUN;

*changes in RSST_accruals;
DATA temp1; SET temp1;
IF MISSING(ivao) THEN ivao=0;
nco=(at-act-ivao)-(lt-lct-dltt);
lag_nco=LAG(nco);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1  THEN ch_nco=nco-lag_nco;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_nco=.;
RUN;
DATA temp1; SET temp1;
IF MISSING(ivst) THEN ivst=0;
IF MISSING(pstk) THEN pstk=0;
fin=(ivst+ivao)-(dltt+dlc+pstk);
lag_fin=LAG(fin);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1  THEN ch_fin=fin-lag_fin;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_fin=.;
RUN;
DATA temp1; SET temp1;
ch_rsst=(ch_wc+ch_nco+ch_fin)*2/(at+lag_at);RUN;

*changes in receivables;
DATA temp1; SET temp1;
lag_rect=LAG(rect);
RUN;
```

```
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1  THEN ch_rec=rect-lag_rect;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_rec=.;
RUN;
DATA temp1; SET temp1;
dch_rec=ch_rec*2/(at+lag_at);
RUN;

*changes in inventories;
DATA temp1; SET temp1;
lag_invt=LAG(invt);
RUN;
DATA temp1;SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN ch_inv=invt-lag_invt;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1  THEN ch_inv=.;
RUN;
DATA temp1; SET temp1;
dch_inv=ch_inv*2/(at+lag_at);RUN;

*percentage of soft assets;
DATA temp1; SET temp1;
soft_assets=(at-ppent-che)/at;
RUN;

*percentage change in cash sales;
DATA temp1; SET temp1;
lag_rect=LAG(rect);RUN;
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN cs=sale-(rect-lag_rect);
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN cs=.;
RUN;
DATA temp1; SET temp1;
lag_cs=LAG(cs);RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1  THEN ch_cs=(cs-lag_cs)/lag_cs;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_cs=.;
RUN;

*change in cash margin;
DATA temp1; SET temp1;
lag_ap=LAG(ap);
lag_invt=LAG(invt);
lag_rect=LAG(rect);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN cmm=(cogs-(invt-lag_invt)+(ap-
lag_ap))/(sale-(rect-lag_rect));
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN cmm=.;
RUN;
DATA temp1;SET temp1;
cm=1-cmm;RUN;
DATA temp1; SET temp1;
lag_cm=LAG(cm);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN ch_cm=(cm-lag_cm)/lag_cm;
```

```
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_cm=.;
RUN;
DATA temp1; SET temp1;
IF MISSING(cogs) THEN ch_cm=.;
IF MISSING (sale) THEN ch_cm=.;
RUN;

*change in return on assets;
DATA temp1; SET temp1;
roa=ni*2/(at+lag_at);
lag_roa=LAG(roa);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN ch_roa=roa - lag_roa;
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN ch_roa=.;
RUN;

*changes in free cash flow;
DATA temp1; SET temp1;
lag_ib=LAG(ib);
RUN;
DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN ch_ib=(ib-
lag_ib)*2/(at+lag_at);
ELSE IF lag_gvkey^=gvkey or lag_fyear^=fyear-1 THEN ch_ib=.;
RUN;

*actual issuance;
DATA temp1;SET temp1;
IF sstk>0 THEN issue=1;
ELSE IF dltis>0 THEN issue=1;
ELSE IF MISSING(sstk) AND MISSING(dltis) THEN issue=.;
ELSE issue=0;
RUN;

*Book-to-market;
DATA temp1; SET temp1;
bm=ceq/(prcc_f*csho);
RUN;


*Depreciation Index (Ratio from Beneish 1999);
DATA temp1; SET temp1;
lag_dp=LAG(dp);
lag_ppent=LAG(ppent);
RUN;

DATA temp1; SET temp1;
IF lag_gvkey=gvkey AND lag_fyear=fyear-1 THEN
dpi=(lag_dp/(lag_dp+lag_ppent))/(dp/(dp+ppent));
ELSE IF lag_gvkey^=gvkey OR lag_fyear^=fyear-1 THEN dpi=.;
RUN;
```

```
*Retained earnings over assets;
*Earnings before interest and tax (Ratios from Summers and Sweeney, 1998);
DATA temp1; SET temp1;
reoa=re/at;
EBIT=(ni+xint+txt)/at;
RUN;

%macro winsor(dsetin=, dsetout=, byvar=none, vars=, type=winsor, pctl=1 99);

%if &dsetout = %then %let dsetout = &dsetin;

%let varL=;
%let varH=;
%let xn=1;

%do %until ( %scan(&vars,&xn)= );
    %let token = %scan(&vars,&xn);
    %let varL = &varL &token.L;
    %let varH = &varH &token.H;
    %let xn=%EVAL(&xn + 1);
%end;

%let xn=%eval(&xn-1);

data xtemp;
    set &dsetin;
    run;

%if &byvar = none %then %do;

    data xtemp;
        set xtemp;
        xbyvar = 1;
        run;

    %let byvar = xbyvar;

%end;

proc sort data = xtemp;
    by &byvar;
    run;

proc univariate data = xtemp noprint;
    by &byvar;
    var &vars;
    output out = xtemp_pctl PCTLPTS = &pctl PCTLPRE = &vars PCTLNAME = L H;
    run;

data &dsetout;
    merge xtemp xtemp_pctl;
    by &byvar;
    array trimvars{&xn} &vars;
    array trimvarl{&xn} &varL;
    array trimvarh{&xn} &varH;

    do xi = 1 to dim(trimvars);
```

```
        %if &type = winsor %then %do;
            if not missing(trimvars{xi}) then do;
              if (trimvars{xi} < trimvarl{xi}) then trimvars{xi} =
trimvarl{xi};
              if (trimvars{xi} > trimvarh{xi}) then trimvars{xi} =
trimvarh{xi};
            end;
        %end;

        %else %do;
            if not missing(trimvars{xi}) then do;
              if (trimvars{xi} < trimvarl{xi}) then delete;
              if (trimvars{xi} > trimvarh{xi}) then delete;
            end;
        %end;

    end;
    drop &varL &varH xbyvar xi;
    run;

%mend winsor;
run;
%winsor(dsetin=temp1, dsetout=temp1, byvar=none, vars=dch_wc ch_rsst dch_rec
dch_inv soft_assets ch_cs ch_cm ch_roa ch_ib
bm dpi reoa ebit, type=winsor, pctl=1 99);
 RUN;

DATA temp1; SET temp1;
ch_fcf=ch_ib-ch_rsst;
RUN;

DATA temp;SET temp1;
KEEP P_AAER gvkey cik sich insbnk fyear datadate understatement misstate
dch_wc ch_rsst dch_rec dch_inv soft_assets
ch_cs ch_cm ch_roa ch_fcf issue bm dpi reoa ebit
at ceq che cogs csho dlc dltt dp ib invt ivao ivst lct lt ni ppegt pstk re
rect sale txp txt xint prcc_f act ap dltis sstk;
RUN;

*Including all accounting raw data as fraud prediction features for
additional analysis;
PROC SQL;
CREATE TABLE temp AS SELECT
a.*, b.*
FROM temp AS a LEFT JOIN fraud.compustatindustrial7815 AS b
ON a.gvkey=b.gvkey
AND a.fyear=b.fyear;
QUIT;
RUN;

PROC SORT DATA=temp nodupkey;
BY gvkey fyear; RUN;

DATA fraud.all; SET temp;
RUN;
```