PRACTIUM 2 – MSDS696-FALL18

Week: 8 – Final Report

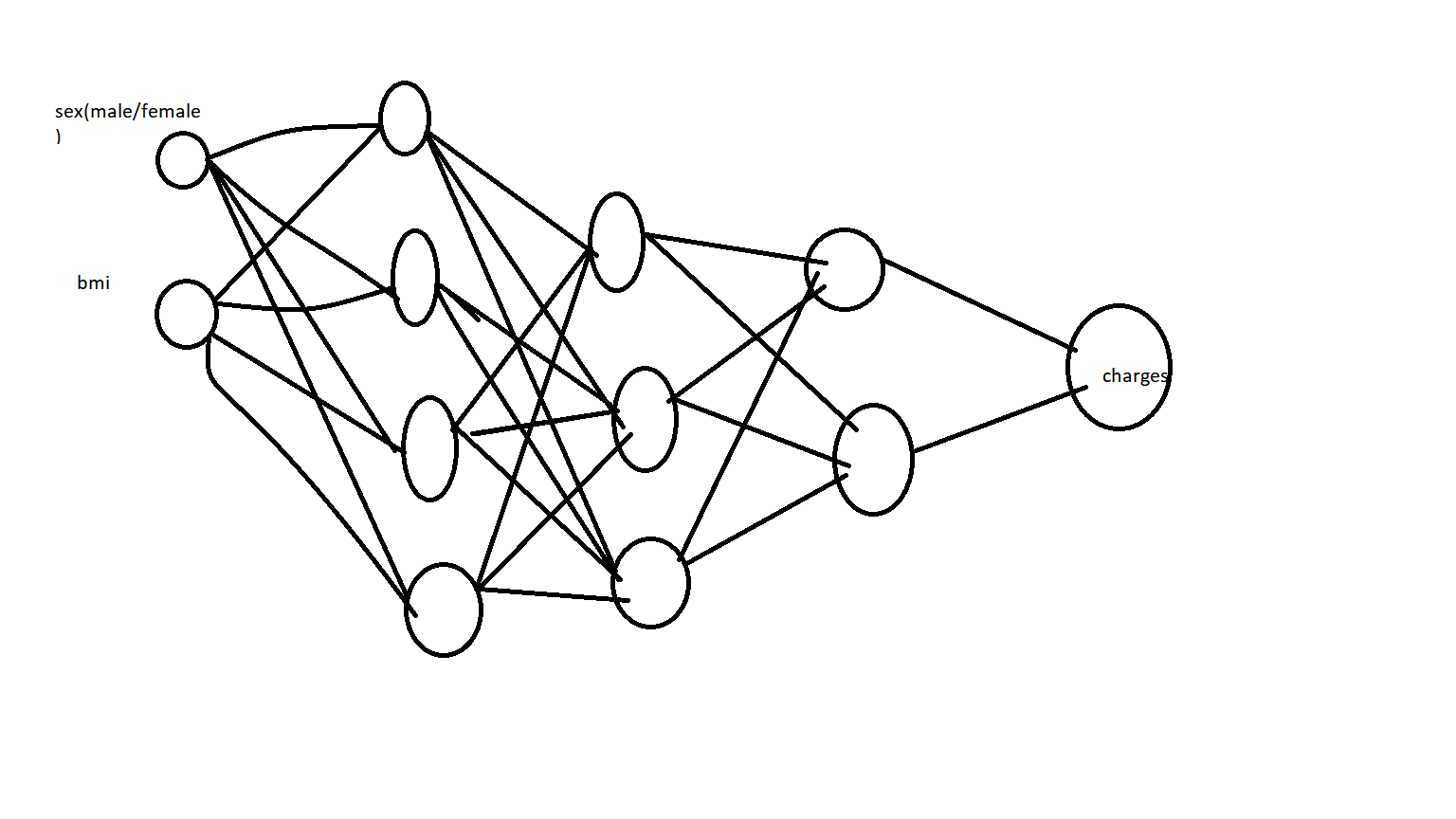
Name: Phong Nguyen

PROJECT TITLE: Deep Learning with TensorFlow: Applied Regression and Neural Network to predict the Health Insurance Cost.

PROJECT SUMMARY: Deep Learning is very subject. In this project, Deep Learning is applied to predict the healthcare insurance price. The technique in Deep Learning is applied TensorFlow. The Methods are Applied Regression and Neural Network.

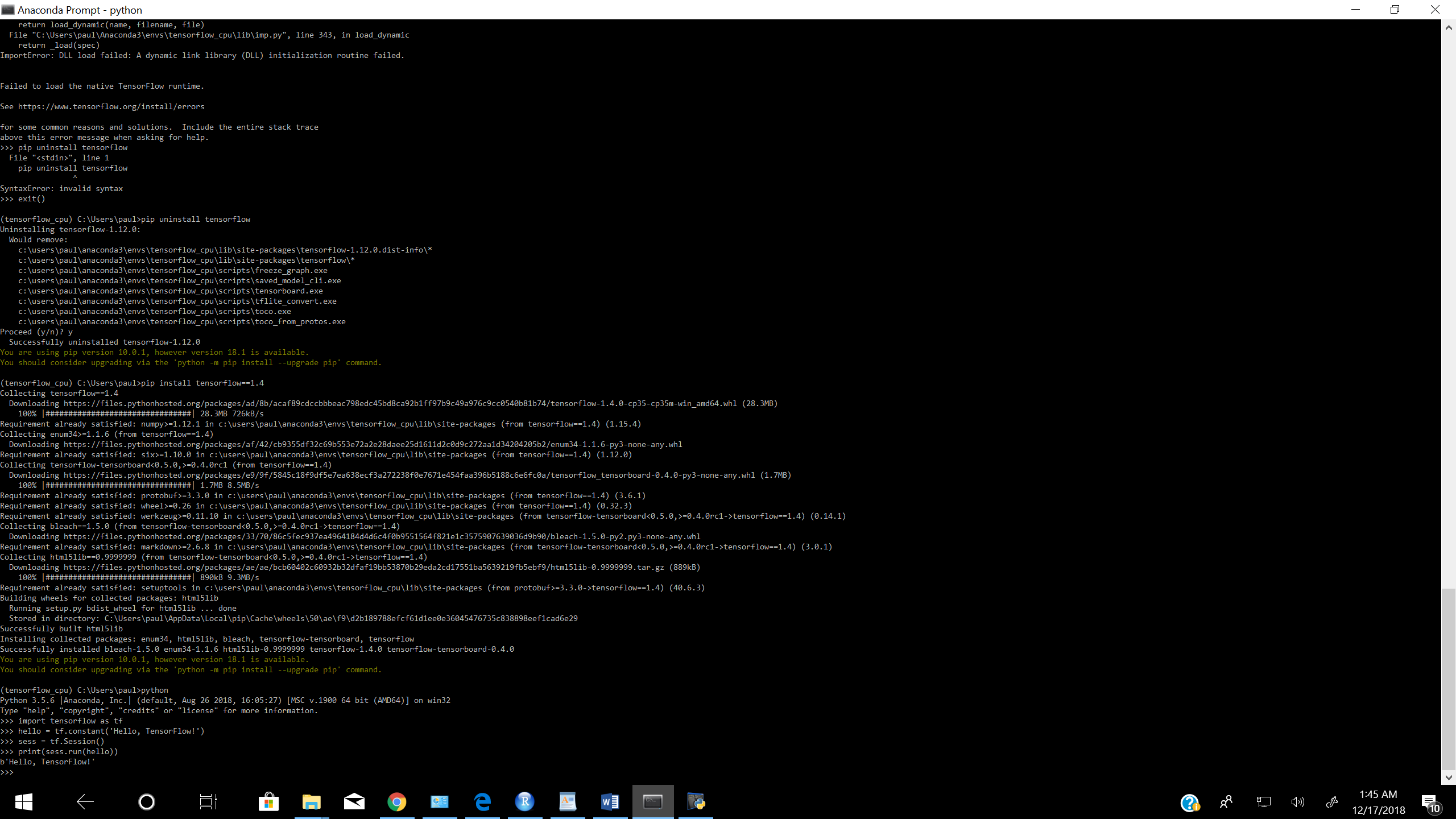
MILESTONES:

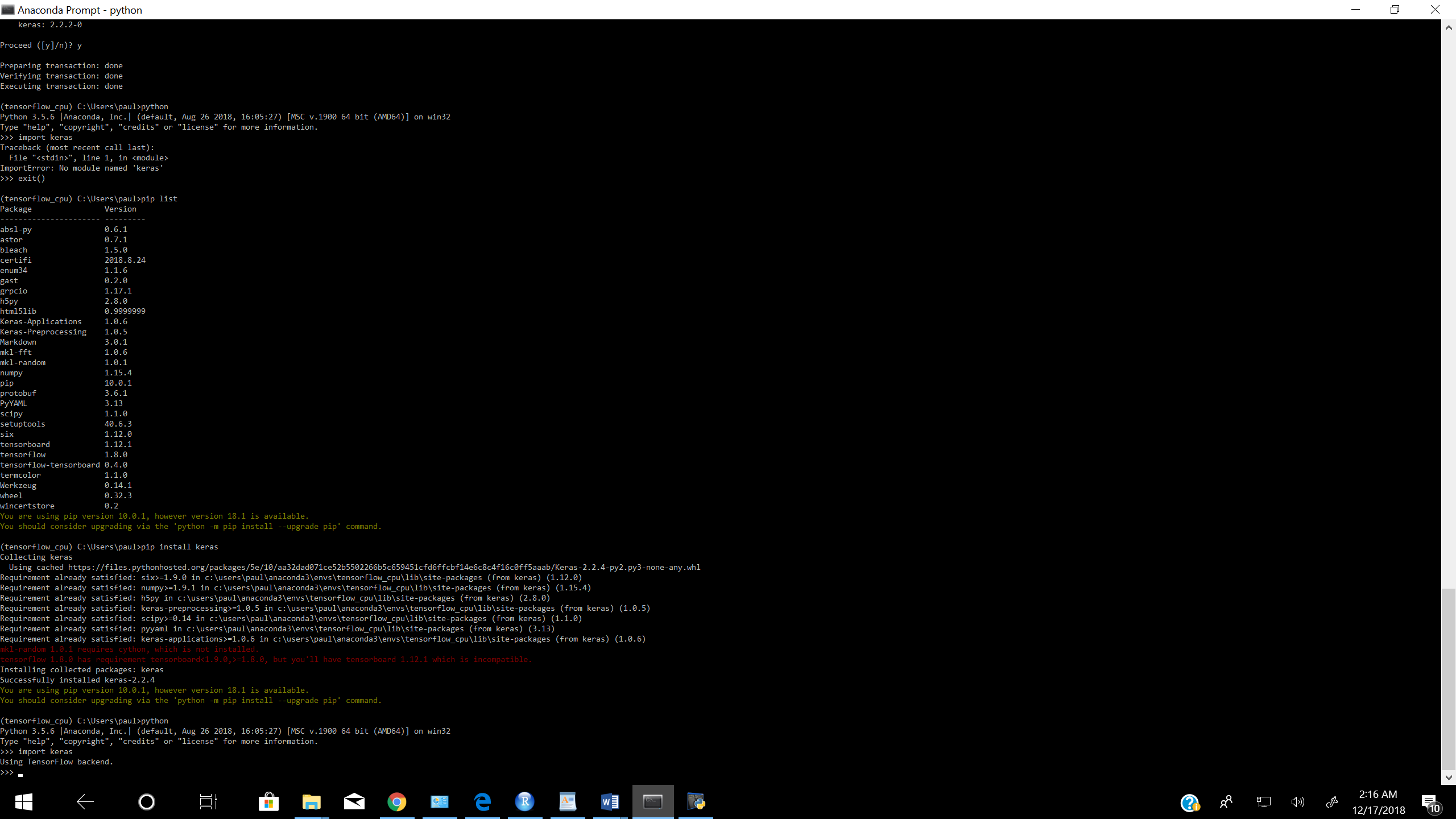
* Overall, collect data and applied some basic analyze to find the relationship between variables in the dataset. There are three variables are used, “charges”, “age” and “bmi” for this project.
* There are two methods are used in this project. It’s Regression and Deep Neural Network. Regression is applied under the frame work of Generalized Additive Models (GAM). In GAM, Lamda-Mu-Sigma is a methods that applied to calculate and predict the results.



Successful created TensorFlow environment and be able to connect between R and Tensor backend.

CREATE TENSORFLOW ENVIROMENT:





This is a big milestone for this project. Because TensorFlow is the main step entitle the project.

EXPLORE THE DATA:

> MFdata=caret::createDataPartition(y=mydata$sex, p=400/600,list=FALSE)

> trainset=mydata[MFdata,]

> testset=mydata[-MFdata,]

> head(trainset)

age sex bmi children smoker region charges

1 19 female 27.90 0 yes southwest 16884.924

3 28 male 33.00 3 no southeast 4449.462

5 32 male 28.88 0 no northwest 3866.855

6 31 female 25.74 0 no southeast 3756.622

7 46 female 33.44 1 no southeast 8240.590

10 60 female 25.84 0 no northwest 28923.137

> trainset%>%gather(charges,key="Outcome",value="Value")%>%

+ ggplot(aes(x=age,y=Value))+

+ geom\_point(aes(col=sex),alpha=0.2)+

+ geom\_smooth(aes(fill=sex,col=sex),alpha=0.5)+

+ theme\_bw()+

+ facet\_grid(Outcome~sex,scales = "free")+

+ scale\_fill\_manual(values=c("red","blue"))+

+ scale\_color\_manual(values=c("red3","blue3"))

`geom\_smooth()` using method = 'loess' and formula 'y ~ x'

> Mdata%>%gather(bmi,age,key="Predictor",value="Value")%>%

+ ggplot(aes(x=Value,y=charges))+

+ geom\_point(shape=21,col="red",fill="red",alpha=0.3)+

+ geom\_smooth(alpha=0.5,fill="red",col="red4")+

+ facet\_wrap(~Predictor,ncol=2,scales="free")+

+ theme\_bw()

`geom\_smooth()` using method = 'loess' and formula 'y ~ x'

> trainset%>%gather(charges,key="Outcome",value="Value")%>%

+ ggplot(aes(x=bmi,y=Value))+

+ geom\_point(aes(col=sex),alpha=0.2)+

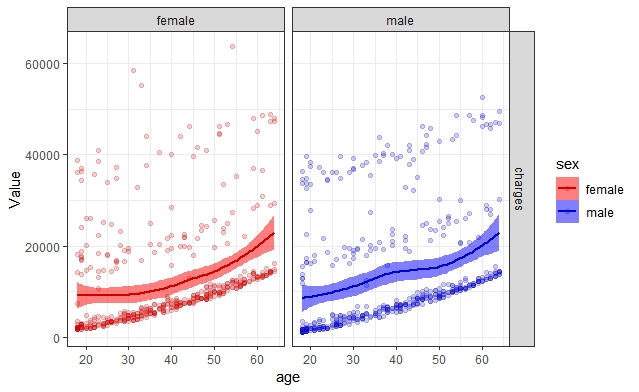
+ geom\_smooth(aes(fill=sex,col=sex),alpha=0.5)+

+ theme\_bw()+

+ facet\_grid(Outcome~sex,scales = "free")+

+ scale\_fill\_manual(values=c("red","blue"))+

+ scale\_color\_manual(values=c("red3","blue3"))



SEPARATE THE DATASET INTO TWO SETS: MALE AND FEMALE.

Create LMS model for GAMLSS [1]

Male model

> numCores<-detectCores()

> MaleModelLMS=gamlss(data=Mtraindata,charges~1,sigma.formula = charges~1,

+ nu.formula = charges~1,family=BCCG(mu.link="log"),trace=FALSE,parallel="multicore",ncpus = numCores)

> MaleModelLMS1=stepGAICAll.A(MaleModelLMS, scope=list(lower=~1,

+ upper=~poly(age,2)+poly(bmi,3)+pb(age)),

+ sigma.scope = list(lower=~1,

+ upper=~poly(age,2)+poly(bmi,3)+pb(age)),

+ k=log(length(Mtraindata)),

+ trace=FALSE,

+ parallel="multicore",ncpus = numCores)

---------------------------------------------------

Start: AIC= 22818.42

charges ~ 1

---------------------------------------------------

Start: AIC= 18656.79

~1

---------------------------------------------------

Start: AIC= 18656.79

~1

---------------------------------------------------

Start: AIC= 18656.79

~1

---------------------------------------------------

Start: AIC= 18656.79

charges ~ poly(bmi, 3) + poly(age, 2) + pb(age)

> summary(MaleModelLMS1)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Family: c("BCCG", "Box-Cox-Cole-Green")

Call: gamlss(formula = charges ~ poly(bmi, 3) + poly(age, 2) + pb(age), sigma.formula = ~1, nu.formula = ~1, family = BCCG(mu.link = "log"),

data = Mtraindata, trace = FALSE, parallel = "multicore", ncpus = numCores)

Fitting method: RS()

------------------------------------------------------------------

Mu link function: log

Mu Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.704e+01 6.063e-02 -610.9 <2e-16 \*\*\*

poly(bmi, 3)1 4.383e-06 1.428e+00 0.0 1

poly(bmi, 3)2 -1.510e-06 1.422e+00 0.0 1

poly(bmi, 3)3 -8.198e-07 1.420e+00 0.0 1

poly(age, 2)1 5.365e-06 1.438e+00 0.0 1

poly(age, 2)2 -5.034e-07 1.417e+00 0.0 1

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------

Sigma link function: log

Sigma Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.038e+01 4.287e-07 70855233 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------

Nu link function: identity

Nu Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.000e-01 4.287e-07 1166190 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------

NOTE: Additive smoothing terms exist in the formulas:

i) Std. Error for smoothers are for the linear effect only.

ii) Std. Error for the linear terms may not be reliable.

------------------------------------------------------------------

No. of observations in the fit: 544

Degrees of Freedom for the fit: 8.000543

Residual Deg. of Freedom: 535.9995

at cycle: 3

Global Deviance: 18642.46

AIC: 18658.46

SBC: 18692.85

FEMALE MODEL:

> FemaleModelLMS1=stepGAICAll.A(FemaleModelLMS, scope=list(lower=~1,

+ upper=~poly(age,2)+poly(bmi,3)+pb(age)),sigma.scope = list(lower=~1, upper=~poly(age,2)+poly(bmi,3)+pb(age)),k=log(length(Ftraindata)),

+ trace=FALSE,parallel="multicore",ncpus = numCores)

---------------------------------------------------

Start: AIC= 16210.46

charges ~ 1

---------------------------------------------------

Start: AIC= 12786.8

~1

---------------------------------------------------

Start: AIC= 12786.8

~1

---------------------------------------------------

Start: AIC= 12786.8

~1

---------------------------------------------------

Start: AIC= 12786.8

charges ~ poly(age, 2) + poly(bmi, 3) + pb(age)

> summary(FemaleModelLMS1)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Family: c("BCCG", "Box-Cox-Cole-Green")

Call: gamlss(formula = charges ~ poly(age, 2) + poly(bmi, 3) + pb(age), sigma.formula = ~1, nu.formula = ~1, family = BCCG(mu.link = "log"),

data = Ftraindata, trace = FALSE, parallel = "multicore", ncpus = numCores)

Fitting method: RS()

------------------------------------------------------------------

Mu link function: log

Mu Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -37.018839 0.061430 -602.623 <2e-16 \*\*\*

poly(age, 2)1 0.170794 1.421494 0.120 0.904

poly(age, 2)2 0.027638 1.421973 0.019 0.985

poly(bmi, 3)1 0.063116 1.420203 0.044 0.965

poly(bmi, 3)2 0.038941 1.416934 0.027 0.978

poly(bmi, 3)3 0.007683 1.420554 0.005 0.996

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------

Sigma link function: log

Sigma Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.529e+01 1.434e-05 1763593 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------

Nu link function: identity

Nu Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.000e-01 4.344e-07 1151086 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------

NOTE: Additive smoothing terms exist in the formulas:

i) Std. Error for smoothers are for the linear effect only.

ii) Std. Error for the linear terms may not be reliable.

------------------------------------------------------------------

No. of observations in the fit: 530

Degrees of Freedom for the fit: 8.000541

Residual Deg. of Freedom: 521.9995

at cycle: 3

Global Deviance: 12772.46

AIC: 12788.46

SBC: 12822.65

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

CREATE DEEP NEURAL NETWROK: Male model

There are two inputs variables, Sex and BMI

> cleaneddata<-recipe(Mtraindata,charges~sex+age)%>%

+ step\_center(all\_predictors())%>%

+ step\_scale(all\_predictors())

Error in eval(predvars, data, env) : object 'sex' not found

> cleaneddata<-recipe(Mtraindata,charges~bmi+age)%>%

+ step\_center(all\_predictors())%>%

+ step\_scale(all\_predictors())

> std\_func<-prep(cleaneddata,training=Mtraindata,retain=F)

> std\_test<-bake(cleaneddata, new\_data=Mtestdata)

Error: At least one step has not been trained. Please run `prep`.

> std\_test<-bake(cleaneddata, new\_data=Mtraindata)

Error: At least one step has not been trained. Please run `prep`.

> std\_test<-bake(cleaneddata, new\_data=Mtestdata)

Error: At least one step has not been trained. Please run `prep`.

> head(Mtraindata)

age bmi children smoker region charges

1 18 33.770 1 no southeast 1725.552

2 28 33.000 3 no southeast 4449.462

3 33 22.705 0 no northwest 21984.471

4 32 28.880 0 no northwest 3866.855

5 37 29.830 2 no northeast 6406.411

6 25 26.220 0 no northeast 2721.321

> std\_test<-bake(std\_func, new\_data=Mtestdata)

> std\_train<-bake(std\_func, new\_data=Mtraindata)

> male\_train\_data<-std\_train%>%

+ dplyr::select(-charges)%>%

+ as.matrix()%>%

+ as.array.default(dimnames=NULL)

> male\_test\_data<-std\_test%>%

+ dplyr::select(-charges)%>%

+ as.matrix()%>%

+ as.array.default(dimnames=NULL)

> male\_train\_targets<-Mtraindata%>%

+ dplyr::select(charges)%>%

+ as.matrix()%>%

+ as.array.default(dimnames=NULL)

> male\_test\_targets<-Mtestdata%>%

+ dplyr::select(charges)%>%

+ as.matrix()%>%

+ as.array.default(dimnames=NULL)

> build\_model <- function() {

+ model <- keras\_model\_sequential() %>%

+ layer\_dense(units = 128, activation = "relu",

+ input\_shape = dim(male\_train\_data)[[2]]) %>%

+ layer\_dense(units = 128, activation = "relu") %>%

+ layer\_dense(units = 128, activation = "relu") %>%

+ layer\_dense(units =128, kernel\_regularizer = regularizer\_l2(0.001))%>%

+ layer\_dense(units = 1)

+

+ model %>% compile(

+ optimizer = "rmsprop",

+ loss = "mae",

+ metric= "msle"

+ )

+ }

> model <- build\_model()

> model %>% fit(male\_train\_data, male\_train\_targets,

+ epochs = 200, batch\_size = 30,verbose=0,

+ validation\_split = 0.1) -> deepneural

> summary(model)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

========================================================================================================================================================================

dense\_1 (Dense) (None, 128) 384

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_2 (Dense) (None, 128) 16512

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_3 (Dense) (None, 128) 16512

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_4 (Dense) (None, 128) 16512

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_5 (Dense) (None, 1) 129

========================================================================================================================================================================

Total params: 50,049

Trainable params: 50,049

Non-trainable params: 0

Female Model:

CONCLUSION:

Reference:

[1] “Generalized Additive Models for Location Scale and Shape (GAMLSS) in R” Rigby RA. and Stasinopoulos DM <https://www.jstatsoft.org/article/view/v023i07/v23i07.pdf>

[2] <https://people.richland.edu/james/lecture/m170/ch08-int.html>