Mini Group Project

4/6/2022

Data Dictionary

In this group project, you will work with analysts' forecast data of earning per share (EPS) provided by Wharton Research Data Services (WRDS). Institutional Brokers' Estimate System (I/B/E/S) provides historical data on certain financial indicators collected from thousands of individual analysts working in more than 3,000 broker houses.

- **Estimator**: Sell-side institution (mostly broker house). It is just the broker.
- **Analyst**: The person who makes the forecast and work for sell-side institution.

Estimators and analysts are represented by codes to hide their real names.

- **TICKER**: A unique identifier assigned to each security. In this group project, you will only model "AAPL" ticker.
- **CNAME**: Company name
- **FPI**: Forecast Period Indicator: The forecasting period. 6: Next Fiscal Quarter 1: Next Fiscal Year
- **FPEDATS**: The Forecast Period End Date: It is the ending date of the fiscal period to which the estimate applies. For the majority of companies, the FPEDATS date is December 31st of that year.
- **ANNDATS**: The Announce date: It is the date on which the analyst first made that particular estimate.
- **ACTDATS**: The Activation date: It is the date when the analyst forecast became effective within the IBES database.
- **REVDATS**: The Review Date: It is the most recent date on which IBES called the analyst and verified that particular estimate as still valid for that analyst. If an analyst confirms that a previous estimate is still valid, the original database record for that estimate is retained and only the REVDATS variable is updated. If an analyst changes their estimate for a given company, a new record is entered in the database with a new ANNDATS. The old record of the analyst (containing the previous estimate) is retained in the database.
- **ANNDATS_ACT**: The Announced date of Actual EPS: The actual EPS value is announced by the company at this date.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("E:/7/martin/min/MiniProject1-1")
IBES<-read.csv("IBES.csv", header=TRUE)
head(IBES,n=1)
```

```
##   TICKER          CNAME  ACTDATS ESTIMATOR ANALYS FPI MEASURE  VALUE  ANNDATS
## 1   AAPL APPLE COMPUTER 19950120       206   7985   6     EPS 0.0095 19950120
##    REVDATS  FPEDATS ANNDATS_ACT ACTUAL
## 1 19950608 19950331    19950420 0.0053
```

- **First row in IBES data set**: On 20-Jan-95 (ANNDATS), analyst 7985 (ANALYS) at Estimator 206 (ESTIMATOR) predicts that the EPS (MEASURE) for Apple Computer (CNAME) with a ticker of AAPL (TICKER) with forecast period ending 31-Mar-95 (FPEDATS) is$0,0095 (VALUE). This estimates was entered into the database on 20-Jan-95 (ACTDATS). On 20-Apr-95(ANNDATS_ACT), APPLE announced an actual EPS of $0.0053 (ACTUAL) for this quarter (FPI).

Know your data

- **Task 1**: Drop rows from IBES data set when a variable has a missing value.

```
IBES <- IBES[complete.cases(IBES),]
dim(IBES)
```

```
## [1] 10674    13
```

- **Task 2**: Drop rows from IBES data set the quarterly forecasts (FPI=6)

```
IBES <- IBES %>% filter(!FPI == 6)
unique(IBES$FPI)
```

```
## [1] 1
```

- **Task 3**: How many analyst provides forecasts for the fiscal year 2021 09-30? Write an R code which calculate that total number and name it **analysts_2021**.

```
IBES <- IBES %>%
  mutate(ACTDATS = as.Date(as.character(ACTDATS),format = "%Y%m%d"),
         ANNDATS = as.Date(as.character(ANNDATS),format = "%Y%m%d"),
         FPEDATS = as.Date(as.character(FPEDATS),format = "%Y%m%d"),
         REVDATS = as.Date(as.character(REVDATS),format = "%Y%m%d"),
         ANNDATS_ACT = as.Date(as.character(ANNDATS_ACT),
                               format = "%Y%m%d"))
IBES$ESTIMATOR <- as.character(IBES$ESTIMATOR)
IBES$ANALYS <- as.character(IBES$ANALYS)
```

```
analysts_2021 <- IBES %>% filter(FPEDATS == "2021-09-30") %>%
  distinct(ANALYS) %>%
  count()

analysts_2021
```

```
##    n
## 1 45
```

- **Task 4**: How many broker house (ESTIMATOR) provides forecasts for the fiscal year 2021 09-30? Write an R code which calculate that total number and name it **brokers_2021**.

```
brokers_2021 <- IBES %>% filter(FPEDATS == "2021-09-30") %>%
  distinct(ESTIMATOR) %>%
  count()
brokers_2021
```

```
##    n
## 1 42
```

We have used distinct to avoid repeating the estimators, we want only unique estimators

- **Task 5**: Which broker house (ESTIMATOR) has the highest number of analysts that provide forecasts for the fiscal year 2021 09-30? Write an R code which calculate that total number and name it **largebroker_2021**.

```
largebroker_2021 <- IBES %>% filter(FPEDATS == "2021-09-30") %>%
  group_by(ESTIMATOR) %>%
  count() %>%
  arrange(desc(n)) %>%
  head(1)
largebroker_2021
```

```
## # A tibble: 1 x 2
## # Groups:   ESTIMATOR [1]
##   ESTIMATOR     n
##   <chr>     <int>
## 1 192          12
```

But we should clearly note that this value will tend to be smaller if we require the analyst to be unique(not repeating in that estimator)

Data filtering and Creation of new variables

- **Task 6**: It is quite possible that an analyst makes multiple forecasts throughout the year for the same fiscal period. Remove observations from the data set if an analyst has multiple predictions for the same year and keep the last one. This step is crucial for successful execution of the following tasks.

```
IBES <- IBES %>%
  group_by(ANALYS, FPEDATS) %>%
  slice(which.max(REVDATS)) %>%
  ungroup()
```

- **Task 7**: **Previous period accuracy**: For each year, calculate the forecast performance of each analyst from last year and name it as **past_accuracy**. In the calculation of forecast performance, you can use the MEASURE-ACTUAL as the forecast accuracy measure.

```
IBES <- IBES %>% mutate(past_accuracy = VALUE - ACTUAL)
names(IBES)
```

```
##  [1] "TICKER"        "CNAME"         "ACTDATS"       "ESTIMATOR"
##  [5] "ANALYS"        "FPI"           "MEASURE"       "VALUE"
##  [9] "ANNDATS"       "REVDATS"       "FPEDATS"       "ANNDATS_ACT"
## [13] "ACTUAL"        "past_accuracy"
```

- **Task 8**: **Forecast Horizon**: The longer the forecast horizon, the higher the uncertainty associated with EPS forecast. To control for this fact, create a new variable called **horizon** that captures the forecast horizon (ANNDATS_ACT- ANNDATS) for each analyst.

```
IBES <- IBES %>% mutate(horizon = ANNDATS_ACT - ANNDATS)
names(IBES)
```

```
##  [1] "TICKER"        "CNAME"         "ACTDATS"       "ESTIMATOR"
##  [5] "ANALYS"        "FPI"           "MEASURE"       "VALUE"
##  [9] "ANNDATS"       "REVDATS"       "FPEDATS"       "ANNDATS_ACT"
## [13] "ACTUAL"        "past_accuracy" "horizon"
```

- **Task 9**: **Experience**: We assume that if an analyst is monitoring a company for a long period of time, he/she is expected to make more informed predictions. Create a new variable called **experience** that counts the cumulative number of years the analyst monitor (have predictions) the company.

```
IBES$mk <- 1
IBES <- IBES %>% group_by(ANALYS) %>%
  mutate(experience = cumsum(mk))

IBES <- select(IBES, -mk)
names(IBES)
```

```
##  [1] "TICKER"        "CNAME"         "ACTDATS"       "ESTIMATOR"
##  [5] "ANALYS"        "FPI"           "MEASURE"       "VALUE"
##  [9] "ANNDATS"       "REVDATS"       "FPEDATS"       "ANNDATS_ACT"
## [13] "ACTUAL"        "past_accuracy" "horizon"       "experience"
```

- **Task 10**: **Brokerage size**: If a brokerage house have many analysts making predictions for the same company, it can be a sign of more resources allocated for company analysis. Create a new variable called **size** that counts the total number of analyst employed per year by the brokerage house (ESTIMATOR).

```
IBES <- IBES %>% group_by(FPEDATS, ESTIMATOR) %>%
  mutate(size = n()) %>%
  ungroup()
names(IBES)
```

```
##  [1] "TICKER"        "CNAME"         "ACTDATS"       "ESTIMATOR"
##  [5] "ANALYS"        "FPI"           "MEASURE"       "VALUE"
##  [9] "ANNDATS"       "REVDATS"       "FPEDATS"       "ANNDATS_ACT"
## [13] "ACTUAL"        "past_accuracy" "horizon"       "experience"
## [17] "size"
```

Modeling

- **Task 11**: **Benchmark Consensus forecast**: As a benchmark, take the average forecasts by all analysts as your best forecast of the EPS for that year and store your forecast of APPL EPS (ACTUAL) for each year in a data frame named as **consensus_forecast**.

```
consensus_forecast <- IBES %>% group_by(FPEDATS) %>%
  summarize(avg = mean(VALUE))
head(consensus_forecast)
```

```
## # A tibble: 6 x 2
##   FPEDATS        avg
##   <date>       <dbl>
## 1 1995-09-30  0.0315
## 2 1996-09-30 -0.0332
## 3 1997-09-30 -0.0246
## 4 1998-09-30  0.0141
## 5 1999-09-30  0.0230
## 6 2000-09-30  0.0311
```

- **Task 12**: **Linear Model**: Come up with a method that uses **past_accuracy**, **horizon**, **experience**, and **size** as features to predict the EPS (**ACTUAL**) with linear regression method. Store your predictions in a data frame named as **OLS_forecast**. Explain your method briefly.

```
price_model <- lm(ACTUAL ~ past_accuracy + horizon + experience + size, data = IBES)

OLS_forecast <- IBES%>% transmute(ACTUAL, predictions = predict(price_model,newdata =  IBES))
head(OLS_forecast)
```

```
## # A tibble: 6 x 2
##    ACTUAL predictions
##     <dbl>       <dbl>
## 1  1.42         0.170
## 2  0.0302       0.827
## 3 -0.0048       0.965
## 4  0.0059       1.08
## 5  0.0036       1.11
## 6  0.0132       1.36
```

```
summary(price_model)
```

```
##
## Call:
## lm(formula = ACTUAL ~ past_accuracy + horizon + experience +
##     size, data = IBES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4296 -0.8971 -0.3415  0.6671  5.8490
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.8042519  0.1373421   5.856 6.27e-09 ***
## past_accuracy -3.5183605  0.3577551  -9.835  < 2e-16 ***
## horizon       -0.0003234  0.0004582  -0.706    0.480
## experience     0.1370723  0.0117551  11.661  < 2e-16 ***
## size          -0.1070575  0.1006811  -1.063    0.288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.271 on 1093 degrees of freedom
## Multiple R-squared:  0.1798, Adjusted R-squared:  0.1768
## F-statistic:  59.9 on 4 and 1093 DF,  p-value: < 2.2e-16
```

Our price model linear regression has the equation: **Actual = -3.5183(past_accuracy) - 0.0003(horizon) + 0.1370(experience) - 0.1071(size) + 0.8043** Our model does not predict well, it has a worse accuracy of about 17.98% and it only explains 17.68% variability in `VALUE`(EPS). It has also a worse residual standard error of 1.271 which implies it will tend to miss most of the values by this value.

- **Task 13**: **Best Subset Model**: Come up with a method that uses any combination of **past_accuracy**, **horizon**, **experience**, and **size** as features to predict the EPS (**ACTUAL**) with linear regression method. Store your predictions in a data frame named as **best_subset_forecast**. Explain your method briefly.

```
best_price_model <- lm(ACTUAL ~ past_accuracy + experience, data = IBES)

best_subset_forecast <- IBES%>% transmute(ACTUAL, predictions = predict(price_model,newdata =  IBES))
head(best_subset_forecast)
```

```
## # A tibble: 6 x 2
##    ACTUAL predictions
##     <dbl>       <dbl>
## 1  1.42         0.170
## 2  0.0302       0.827
## 3 -0.0048       0.965
## 4  0.0059       1.08
## 5  0.0036       1.11
## 6  0.0132       1.36
```

```
summary(best_price_model)
```

```
##
## Call:
## lm(formula = ACTUAL ~ past_accuracy + experience, data = IBES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4303 -0.8897 -0.3187  0.6875  5.8608
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.64291    0.06241  10.301   <2e-16 ***
## past_accuracy -3.44270    0.35152  -9.794   <2e-16 ***
## experience     0.13906    0.01154  12.054   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.271 on 1095 degrees of freedom
## Multiple R-squared:  0.1782, Adjusted R-squared:  0.1767
## F-statistic: 118.7 on 2 and 1095 DF,  p-value: < 2.2e-16
```

Our best price model linear regression has the equation: **Actual = -3.4427(past_accuracy) - 0.0003(horizon) + 0.1391(experience)** Our model does not predict well, it has a worse accuracy of about 17.82% and it only explains 17.67% variability in `VALUE`. It has also a worse residual standard error of 1.271 which implies it will tend to miss most of the values by this value.

- **Task 14**: Write a short summary of your findings. In your summary report, also indicate which model does a great job to fit the data and why? From the coefficients of our two models we can note keenly that, past accuracy tends to reduce the change in value by -3.524 to -3.44 units whenever it tends to increase by 1 unit. We can therefore conclude that past accuracy generally tends to negatively influence EPS.

The EPS tends to be positively influenced by experience, which implies as experience tends to increase then the price prediction of EPS tends to improve. The coefficient of size show clearly a decline in EPS whenever the size tends to be large. This may be due to conflicting predictions from different analysts of the estimator. Horizon tends to influence the prediction negatively, as it increases EPS prediction tends to decline.

We can generally conclude that the price model tends to predict better as compared to best price model even though it is by a smaller margin. It has a higher accuracy and a higher variability as compared to the best price model.