1 model development

GLM model with Poisson Distribution & Hurdle Model with Binomial Distribution

Ing. Peter Tomko, M.A.

1/9/2020

```
con <- DBI::dbConnect(odbc::odbc(), "betting_ds", bigint = "integer")
```

```
select * from t_match_stats;
```

Model Estimation

Two approaches are compared in this place, namely GLM model with Poisson Distribution and hurdle model. The latter one is set with logit link function and binomial distribution - i.e. firstly the logit is estimated to predict if the the team scores and if the team scores the second part consists of Poisson distribution. It means that the special distribution is concerned when the team does not score.

```
model_vars <- c(unique(var_importance$data$Variable), "is_home", "n_goals")
master_data <- master_data %>%
  group_by(data_type) %>%
  mutate(glm_data =
           map(binned_data,
               function(i_data){
                 i_data %>%
                   select(one_of(model_vars)) %>%
                   mutate(is_home = ifelse(is_home == 1, "yes", "no")) %>%
                   rename_all(~stringr::str_replace_all(., "__", "_")) %>%
                   rename_all(~stringr::str_replace_all(., "woe.", "")) %>%
                   rename_all(~stringr::str_replace_all(., ".binned", "")) %>%
                   as.data.frame()
               }))

# - calculate full model
glm_model <-
  glm(n_goals ~ .,
      family = "poisson",
      data = master_data %>%
        filter(data_type %in% "Train") %>%
        select(data_type, glm_data) %>%
        unnest(c(glm_data)) %>%
        as.data.frame() %>%
        select(-data_type))

# - select only subset of variables based on VIF
vif_glm_model <- data.frame("vif" = car::vif(glm_model))
final_vars <- rownames(vif_glm_model %>% filter(vif <= 5))
model_data <-
```

```r
  master_data %>%
  filter(data_type %in% "Train") %>%
  select(data_type, glm_data) %>%
  unnest(c(glm_data)) %>%
  as.data.frame() %>%
  select(-data_type) %>%
  select(n_goals, one_of(final_vars))

# - estimate final models
glm_m_final <- glm(n_goals ~ ., family = "poisson", data = model_data)

hurdle_model <- hurdle(n_goals ~ ., dist = "poisson",
                       link = "logit", zero.dist = "binomial",
                       data = model_data)
```

Summary of GLM Model

```
##
## Call:
## glm(formula = n_goals ~ ., family = "poisson", data = model_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2279  -1.4170  -0.2295   0.5714   4.7570
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.773e-01  3.762e-03  47.117  < 2e-16 ***
## team                    -2.746e-03  7.762e-05 -35.375  < 2e-16 ***
## r_ah_advantage_last_20  -1.328e-03  2.762e-04  -4.809 1.52e-06 ***
## avg_total_goals_last_20 -2.113e-03  1.343e-04 -15.732  < 2e-16 ***
## r_draw_odds_last_10     -1.007e-03  1.872e-04  -5.381 7.39e-08 ***
## league                   8.431e-05  1.963e-04   0.430  0.66751
## r_ah_advantage_last_30  -8.990e-04  2.808e-04  -3.201  0.00137 **
## r_team_odds_last_40     -2.621e-03  7.589e-05 -34.532  < 2e-16 ***
## is_homeyes               2.187e-01  5.100e-03  42.877  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 156672  on 126247  degrees of freedom
## Residual deviance: 149072  on 126239  degrees of freedom
## AIC: 369964
##
## Number of Fisher Scoring iterations: 5
```

Summary of Hurdle Model

```
##
## Call:
## hurdle(formula = n_goals ~ ., data = model_data, dist = "poisson", zero.dist = "binomial",
##     link = "logit")
##
```

```
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.5513 -0.9943 -0.2258  0.6135  7.3461
##
## Count model coefficients (truncated poisson with log link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.036e-01  5.379e-03  37.861  < 2e-16 ***
## team                    -2.884e-03  1.066e-04 -27.058  < 2e-16 ***
## r_ah_advantage_last_20  -1.252e-03  4.189e-04  -2.987  0.00281 **
## avg_total_goals_last_20 -2.229e-03  1.815e-04 -12.283  < 2e-16 ***
## r_draw_odds_last_10     -1.241e-03  2.662e-04  -4.663 3.12e-06 ***
## league                  -7.436e-04  2.694e-04  -2.760  0.00577 **
## r_ah_advantage_last_30  -9.977e-04  4.247e-04  -2.349  0.01882 *
## r_team_odds_last_40     -2.502e-03  9.071e-05 -27.586  < 2e-16 ***
## is_homeyes               1.948e-01  7.056e-03  27.610  < 2e-16 ***
## Zero hurdle model coefficients (binomial with logit link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.8036207  0.0090674  88.627  < 2e-16 ***
## team                    -0.0045097  0.0002050 -22.004  < 2e-16 ***
## r_ah_advantage_last_20  -0.0020946  0.0005905  -3.547 0.000389 ***
## avg_total_goals_last_20 -0.0034764  0.0003569  -9.742  < 2e-16 ***
## r_draw_odds_last_10     -0.0014253  0.0004574  -3.116 0.001832 **
## league                   0.0018671  0.0005062   3.688 0.000226 ***
## r_ah_advantage_last_30  -0.0012812  0.0006028  -2.125 0.033561 *
## r_team_odds_last_40     -0.0060128  0.0003360 -17.898  < 2e-16 ***
## is_homeyes               0.4296794  0.0131627  32.644  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 13
## Log-likelihood: -1.849e+05 on 18 Df
```