Predicting benefit receipt - 3 methods

Using K-nearest neighbours algorithms

Contents


```
model_data <- ukmod_tidy |>
  group_by(year, idhh) |>
  mutate(lba_income = sum(lba_income),
         uc_income = max(uc_income),
         uc_receipt = max(uc_receipt),
         n_hh_emp = sum(employment == "Employed"),
         n_hh_unemp = sum(employment == "Unemployed"),
         n_hh_inact = sum(employment == "Inactive")) |>
  ungroup() |>
  filter(age > 17 & age < 66) |>
  select(-idhh, -i_0, -i_m, -i_l, - income, -employment) |>
  mutate(
    year = as.integer(year),
    # p_hh_emp = if_else(employment == "Employed" & n_hh_emp > 0, 1, 0),
    n_hh_emp = fct_other(factor(n_hh_emp), c("0", "1"), other_level = "2+"),
    n_hh_unemp = fct_other(factor(n_hh_unemp), c("0", "1"), other_level = "2+"),
    n_hh_inact = fct_other(factor(n_hh_inact), c("0", "1"), other_level = "2+"),
    benefit_change = uc_income - lba_income,
    ) |>
  fastDummies::dummy_cols(remove_first_dummy = TRUE, remove_selected_columns = TRUE) |>
  janitor::clean_names() |>
  select(-starts_with("n_hh")) |>
  mutate(uc_receipt = factor(uc_receipt, levels = 1:0, labels = c("Yes", "No")))

set.seed(123)

data_split <- initial_split(model_data, prop = 0.8, strata = year)

train_data <- training(data_split) |> select(-year)
test_data <- testing(data_split) |> select(-year)

knitr::kable(head(model_data[,1:5], 10))
```

year	uc_income	lba_income	uc_receipt	age
2014	233.81	344.76	Yes	50
2014	233.81	344.76	Yes	40
2014	0.00	0.00	No	45

year	uc_income	lba_income	uc_receipt	age
2014	0.00	0.00	No	42
2014	0.00	0.00	No	44
2014	0.00	0.00	No	36
2014	0.00	0.00	No	61
2014	0.00	0.00	No	60
2014	0.00	0.00	No	43
2014	0.00	0.00	No	55

Predicting by individual observation (no household data)

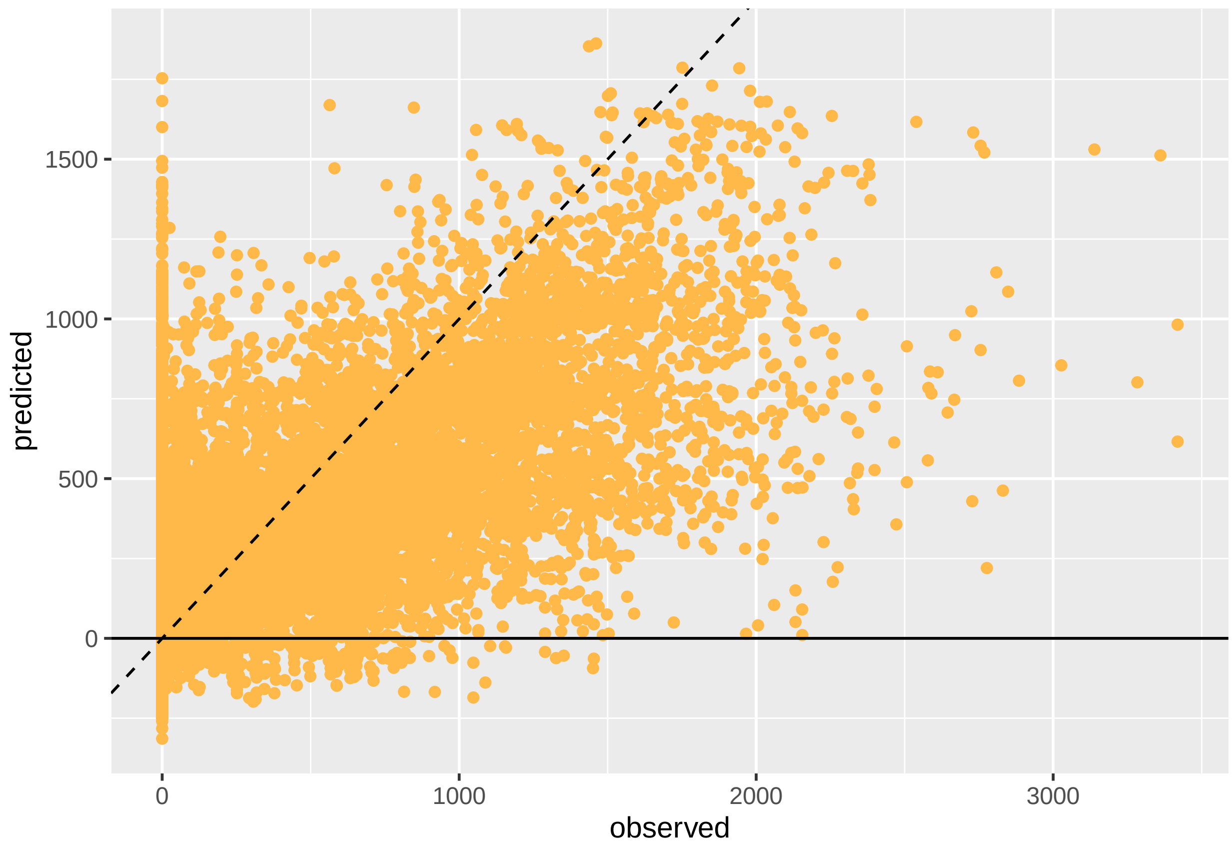
Predicting benefit types separately

UC receipt amount

```
train_data_uc_income <- train_data |> select(-uc_receipt, -lba_income, -benefit_change)
test_data_uc_income <- test_data |> select(-uc_receipt, -lba_income, -benefit_change)

rc_income <- recipe(uc_income ~ .,
                    data = train_data_uc_income) |>
  step_interact(
    ~ starts_with('gender_'):starts_with('children_') +
      starts_with('gender_'):starts_with('children_'):starts_with('emp_len_') +
      starts_with('children_'):starts_with('emp_len_') +
      student:starts_with('children_') + student:starts_with('caring_') +
      starts_with('marsta_') * starts_with('gender_') * starts_with('children_')
  )

mod_ln <- linear_reg(mode = "regression")
mod_knn <- nearest_neighbor(mode = "regression")

wf_ln <- workflow() |>
  add_recipe(rc_income) |>
  add_model(mod_ln)

mod_ln_uc <- fit(wf_ln, data = train_data_uc_income)

pred_uc_income <-
  test_data_uc_income |>
  bind_cols(predict(mod_ln_uc, new_data = test_data_uc_income))

uc_rsq <- rsq_vec(pred_uc_income$uc_income, pred_uc_income$.pred)
uc_rmse <- rmse_vec(pred_uc_income$uc_income, pred_uc_income$.pred)

pred_uc_income |>
  select(observed  = uc_income, predicted = .pred) |>
  ggplot(aes(observed, predicted)) +
  geom_point(colour = sphsu_cols("Pumpkin"), names = FALSE) +
   geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  geom_hline(yintercept = 0) +
  scale_colour_sphsu()
```

```
## Warning in geom_point(colour = sphsu_cols("Pumpkin"), names = FALSE): Ignoring
```

```
## unknown parameters: `names`
```

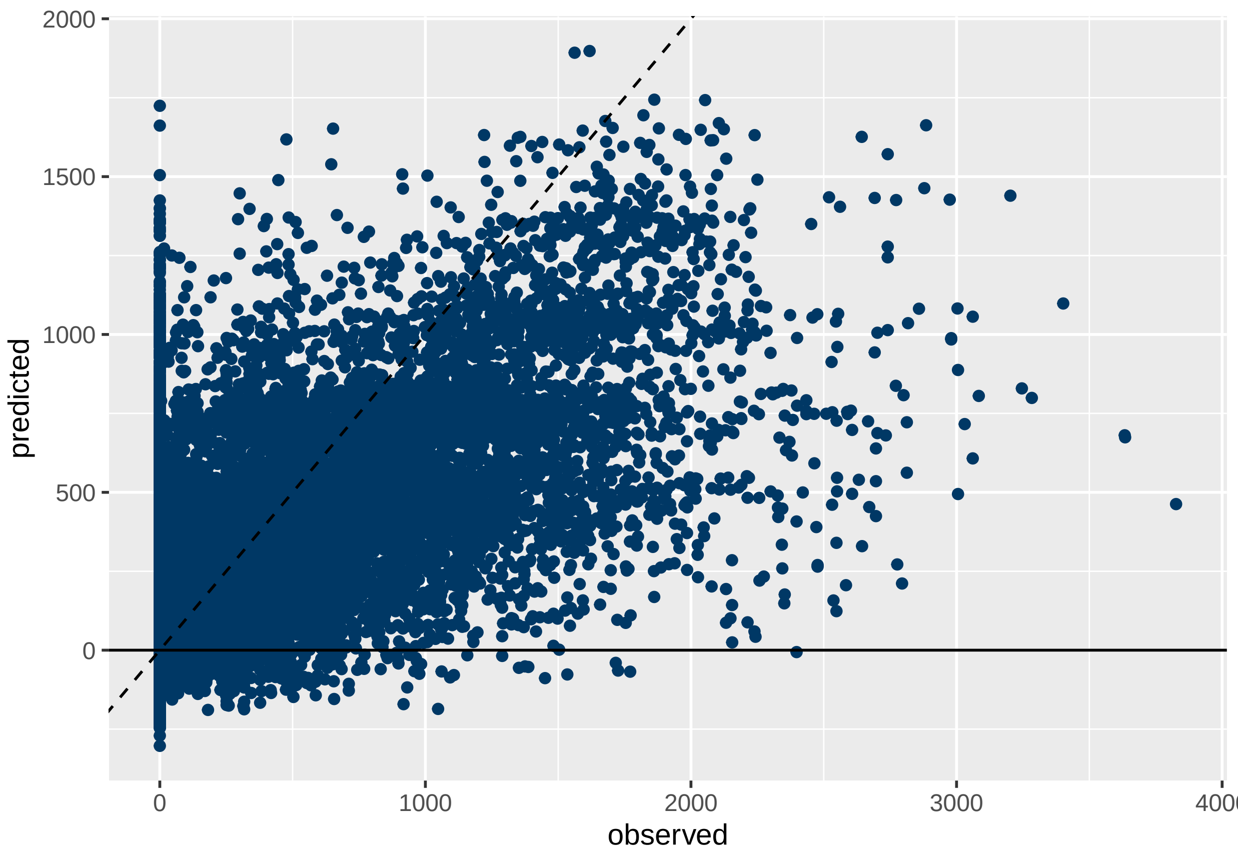


For this model, $R^2 = 0.475$ and root mean squared error $RMSE = 313.6$

Legacy benefit receipt amount

```
train_data_lba_income <- train_data |> select(-uc_receipt, -uc_income, -benefit_change)
test_data_lba_income <- test_data |> select(-uc_receipt, -uc_income, -benefit_change)

rc_income <- recipe(lba_income ~ .,
                 data = train_data_lba_income) |>
  step_interact(
    ~ starts_with('gender_'):starts_with('children_') +
      starts_with('gender_'):starts_with('children_'):starts_with('emp_len_') +
      starts_with('children_'):starts_with('emp_len_') +
      student:starts_with('children_') + student:starts_with('caring_') +
      starts_with('marsta_') * starts_with('gender_') * starts_with('children_')
  )

wf_ln <- workflow() |>
  add_recipe(rc_income) |>
  add_model(mod_ln)

mod_ln_lb <- fit(wf_ln, data = train_data_lba_income)

pred_lb_income <-
```

```r
  test_data_lba_income |>
  bind_cols(predict(mod_ln_lb, new_data = test_data_lba_income))

lb_rsq <- rsq_vec(pred_lb_income$lba_income, pred_lb_income$.pred)
lb_rmse <- rmse_vec(pred_lb_income$lba_income, pred_lb_income$.pred)

pred_lb_income |>
  select(observed  = lba_income, predicted = .pred) |>
  ggplot(aes(observed, predicted)) +
  geom_point(colour = sphsu_cols("University Blue", names = FALSE)) +
   geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  geom_hline(yintercept = 0) +
  scale_colour_sphsu()
```

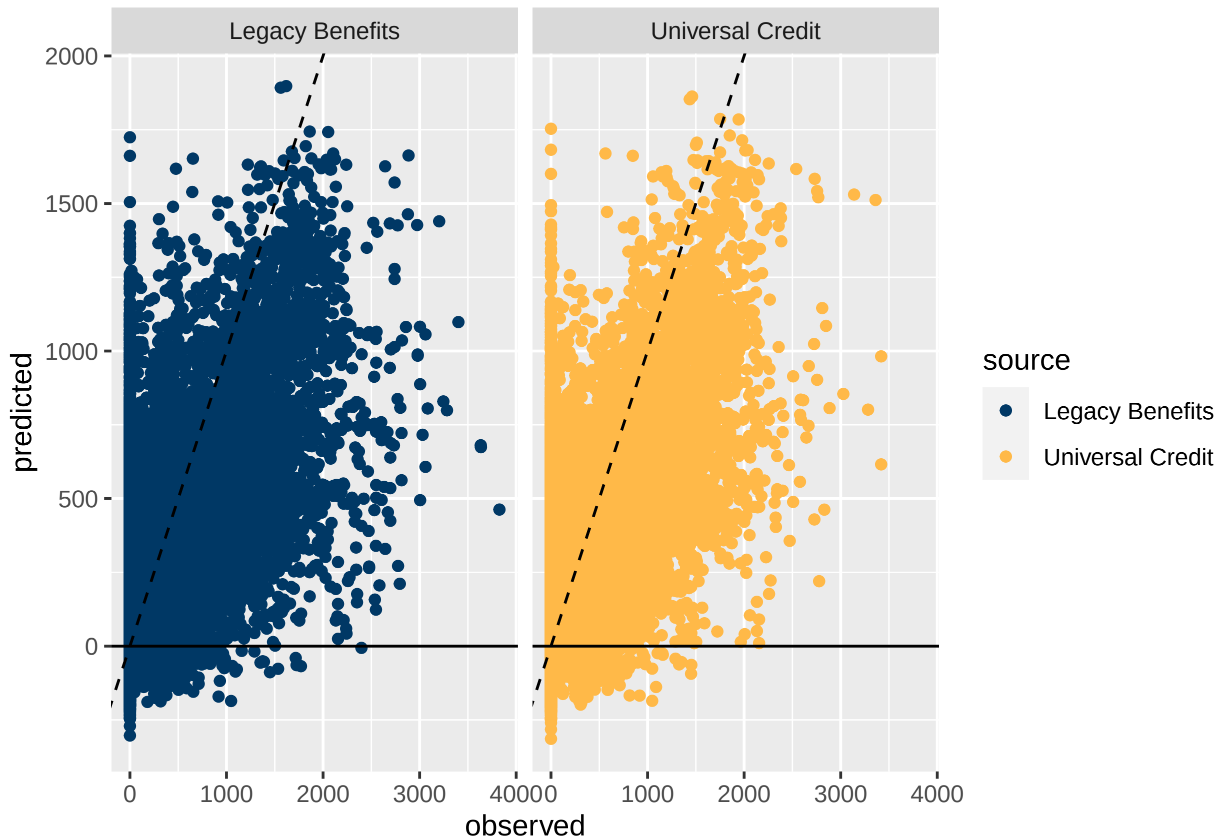


For this model, $R^2 = 0.437$ and root mean squared error $RMSE = 342.1$

Plotting both income types

```r
pred_uc_income |> select(observed = uc_income, predicted = .pred) |>
  mutate(source = "Universal Credit") |>
  bind_rows(
    pred_lb_income |> select(observed = lba_income, predicted = .pred) |>
  mutate(source = "Legacy Benefits")
  ) |>
  ggplot(aes(observed, predicted, colour = source)) +
  geom_point() +
```

```
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  geom_hline(yintercept = 0) +
  facet_wrap(~source) +
  scale_colour_sphsu()
```



Predicting difference

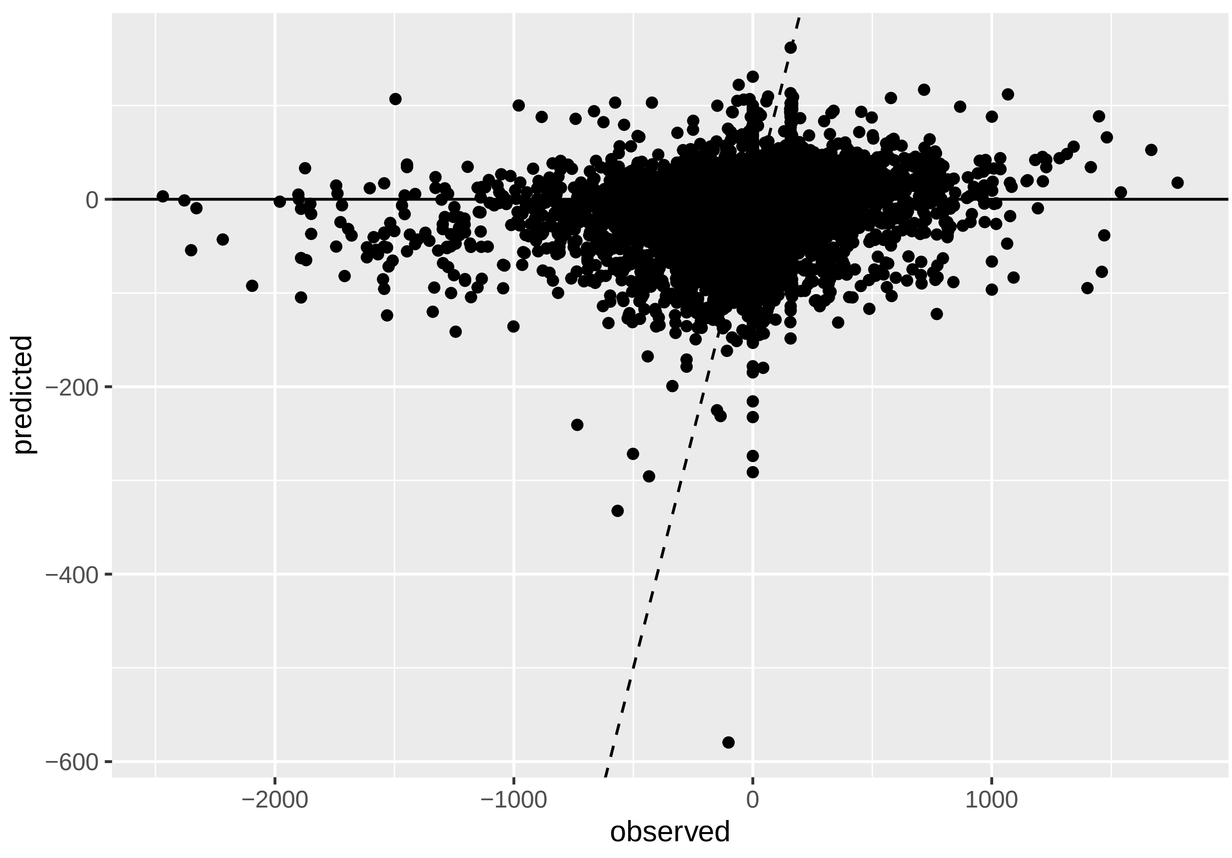
```
train_data_benefit_change <- train_data |> select(-uc_receipt, -uc_income, -lba_income)
test_data_benefit_change <- test_data |> select(-uc_receipt, -uc_income, -lba_income)


rc_income <- recipe(benefit_change ~ .,
                    data = train_data_benefit_change) |>
  step_interact(
    ~ starts_with('gender_'):starts_with('children_') +
      starts_with('gender_'):starts_with('children_'):starts_with('emp_len_') +
      starts_with('children_'):starts_with('emp_len_') +
      student:starts_with('children_') + student:starts_with('caring_') +
      starts_with('marsta_') * starts_with('gender_') * starts_with('children_')
  )

wf_ln <- workflow() |>
  add_recipe(rc_income) |>
  add_model(mod_ln)
```

```
mod_ln_lb <- fit(wf_ln, data = train_data_benefit_change)

pred_benefit_change <-
  test_data_benefit_change |>
  bind_cols(predict(mod_ln_lb, new_data = test_data_benefit_change))

bc_rsq <- rsq_vec(pred_benefit_change$benefit_change, pred_benefit_change$.pred)
bc_rmse <- rmse_vec(pred_benefit_change$benefit_change, pred_benefit_change$.pred)
```

For this model, $R^2 = 0.023$ and root mean squared error $RMSE = 169.1$

Plotting

```
pred_benefit_change |>
  select(observed  = benefit_change, predicted = .pred) |>
  ggplot(aes(observed, predicted)) +
  geom_point() +
   geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  geom_hline(yintercept = 0) +
  scale_colour_sphsu()
```



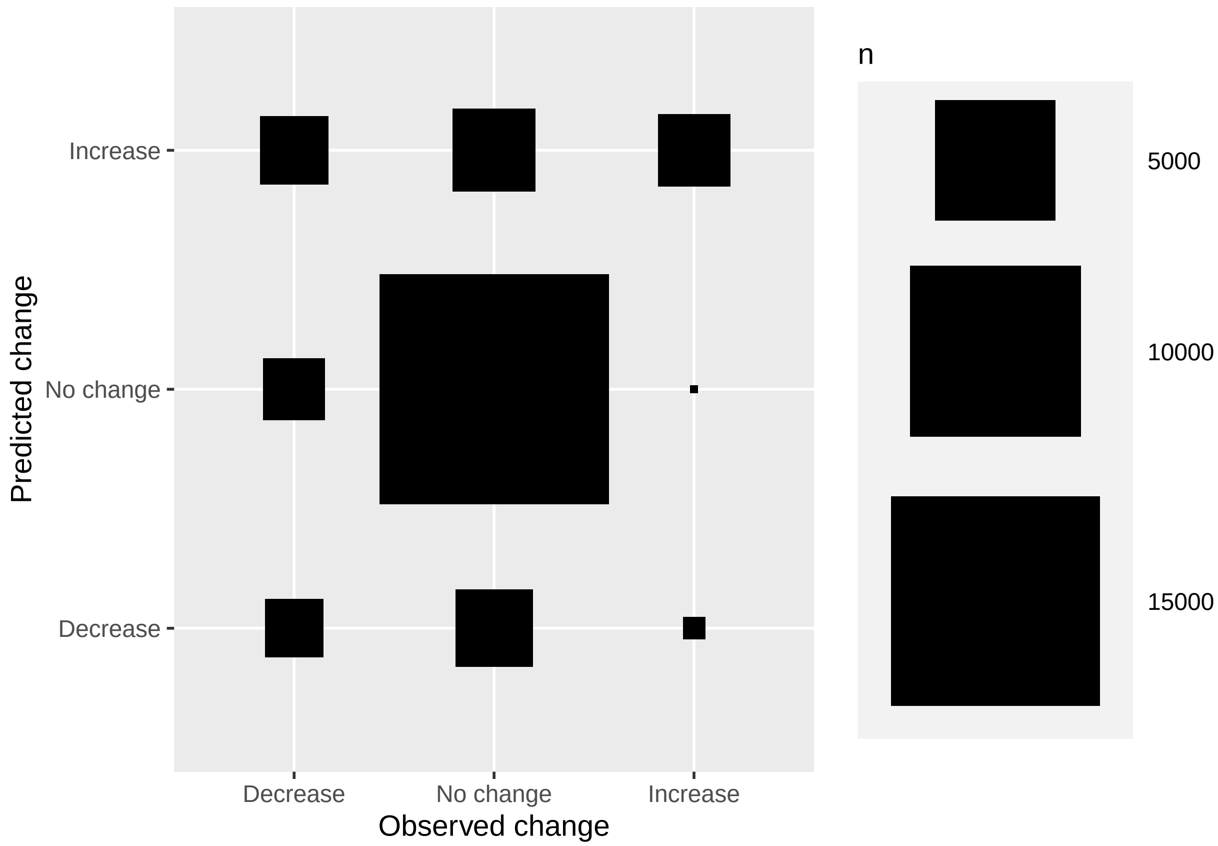
Trichotomise

```
model_data_tri <- model_data |>
  mutate(benefit_change = case_when(
```

```r
    benefit_change == 0 ~ "No change",
    lba_income == 0 & uc_income != 0 ~ "Decrease",
    benefit_change/lba_income >= 0.02 ~ "Increase",
    abs(benefit_change/lba_income) < 0.02 ~ "No change",
    TRUE ~ "Decrease"
  ),
  benefit_change = factor(benefit_change, levels = c("Decrease", "No change", "Increase"))) |>
  select(-uc_income, -lba_income, -uc_receipt)

data_split <- initial_split(model_data_tri, prop = 0.8, strata = year)

train_data <- training(data_split) |> select(-year)
test_data <- testing(data_split) |> select(-year)


rc_income <- recipe(benefit_change ~ .,
                 data = train_data) |>
  step_interact(
    ~ starts_with('gender_'):starts_with('children_') +
      starts_with('gender_'):starts_with('children_'):starts_with('emp_len_') +
      starts_with('children_'):starts_with('emp_len_') +
      student:starts_with('children_') + student:starts_with('caring_') +
      starts_with('marsta_') * starts_with('gender_') * starts_with('children_')
  )

mod_mlogit <- multinom_reg(penalty = double(1), mixture = double(1)) |>
  set_engine("glmnet")
mod_knn <- nearest_neighbor(mode = "classification")


imbal_rec <- rc_income |>
  step_smote(benefit_change, over_ratio = 0.75)

wf_ln <- workflow() |>
  add_recipe(imbal_rec) |>
  add_model(mod_mlogit)

mod_ml_tri <- fit(wf_ln, data = train_data)

pred_benefit_change <-
  test_data |>
  bind_cols(predict(mod_ml_tri, new_data = test_data))

pred_benefit_change_prob <-
  test_data |>
  bind_cols(predict(mod_ml_tri, new_data = test_data, type = "prob"))

acc_bc <- pred_benefit_change %$% accuracy_vec(benefit_change, .pred_class)
roc_inc <- pred_benefit_change_prob %$% roc_auc_vec(benefit_change, as.matrix(data.frame(.pred_Decrease

sens_dec <- pred_benefit_change |>
  mutate(ob_decrease = fct_other(benefit_change, keep = "Decrease"),
         pred_decrease = fct_other(.pred_class, keep = "Decrease")) %$%
```

```
    sens_vec(ob_decrease, pred_decrease)

sens_inc <- pred_benefit_change |>
  mutate(ob_increase = fct_other(benefit_change, keep = "Increase"),
         pred_increase = fct_other(.pred_class, keep = "Increase")) %$%
    sens_vec(ob_increase, pred_increase)
```

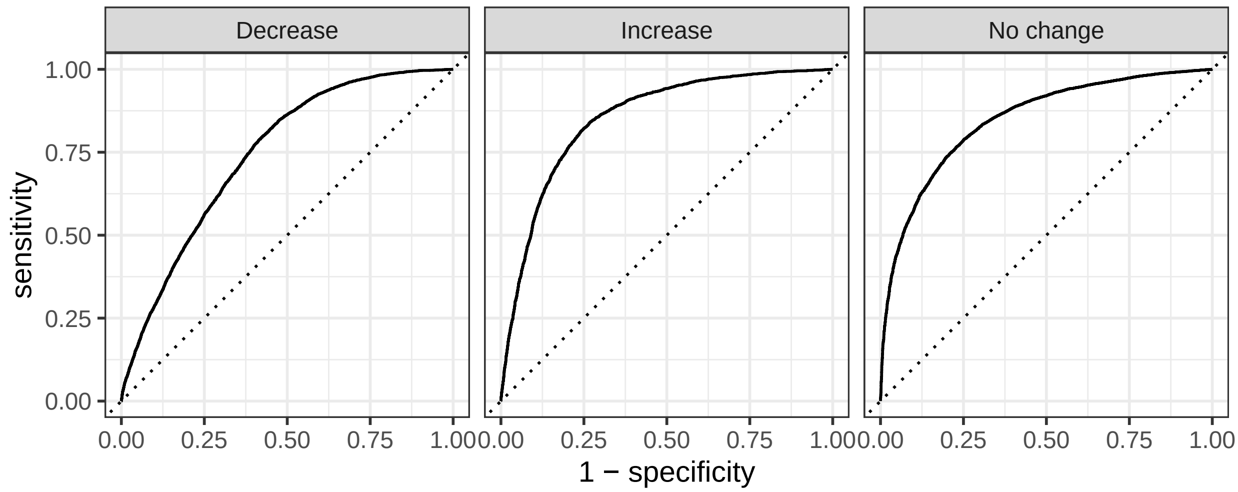
```
pred_benefit_change |>
  ggplot(aes(benefit_change, .pred_class)) +
  geom_count(shape = 15) +
  ylab("Predicted change") +
  xlab("Observed change") +
  scale_size_continuous(range = c(1, 40))
```



This model has an accuracy of 70.8%, a sensitivity for detecting decreasing benefit changes of 29.4% and a sensitivity of detecting increasing benefits of 67.0% (area under ROC-curve = 0.767).

ROC curves

```
pred_benefit_change_prob |> roc_curve(benefit_change, .pred_Decrease, `.pred_No change`, .pred_Increase)
```



Proportions of predictions within each category of observed benefit change:

```
pred_benefit_change %$% table(benefit_change, .pred_class) |> prop.table(margin = 1)
```

```
##                 .pred_class
## benefit_change   Decrease No change  Increase
##      Decrease  0.2943997 0.3240357 0.3815646
##      No change 0.0952904 0.7958440 0.1088657
##      Increase  0.1783859 0.1517146 0.6698995
```