Approach_Zindi_UmojaHack India Income Prediction Challenge

Create a machine learning model to predict whether an individual earns above 50,000 in a specific currency or not.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
  - Replace the numerical column's unknown values.
  - Clean the categorical column's unknown values.
  - Missing value mode imputation for the categorical columns,
    - class
    - occupation_code_main
    - is_hispanic
    - country_of_birth_own
    - country_of_birth_father
    - country_of_birth_mother
    - migration_code_change_in_msa
    - migration_prev_sunbelt
    - migration_code_move_within_reg
    - migration_code_change_in_reg

- Missing value mean imputation for the numerical columns,
    - age
    - wage_per_hour
    - gains
    - losses
    - stocks_status
    - importance_of_record
- Feature Engineering
    - Age check
    - Group by numerical summary
- Missing value indicator

* The final features for the model
  - 1_age
  - 2_gender
  - 3_education
  - 4_class
  - 5_marital_status
  - 6_race
  - 7_is_hispanic
  - 8_employment_commitment
  - 9_employment_stat
  - 10_wage_per_hour
  - 11_working_week_per_year

- 12_industry_code
- 13_industry_code_main
- 14_occupation_code
- 15_occupation_code_main
- 16_total_employed
- 17_household_stat
- 18_household_summary
- 19_vet_benefit
- 20_tax_status
- 21_gains
- 22_losses
- 23_stocks_status
- 24_citizenship
- 25_mig_year
- 26_country_of_birth_own
- 27_country_of_birth_father
- 28_country_of_birth_mother
- 29_migration_code_change_in_msa
- 30_migration_prev_sunbelt
- 31_migration_code_move_within_reg
- 32_migration_code_change_in_reg
- 33_residence_1_year_ago
- 34_importance_of_record
- 35_income_above_limit
- 36_data

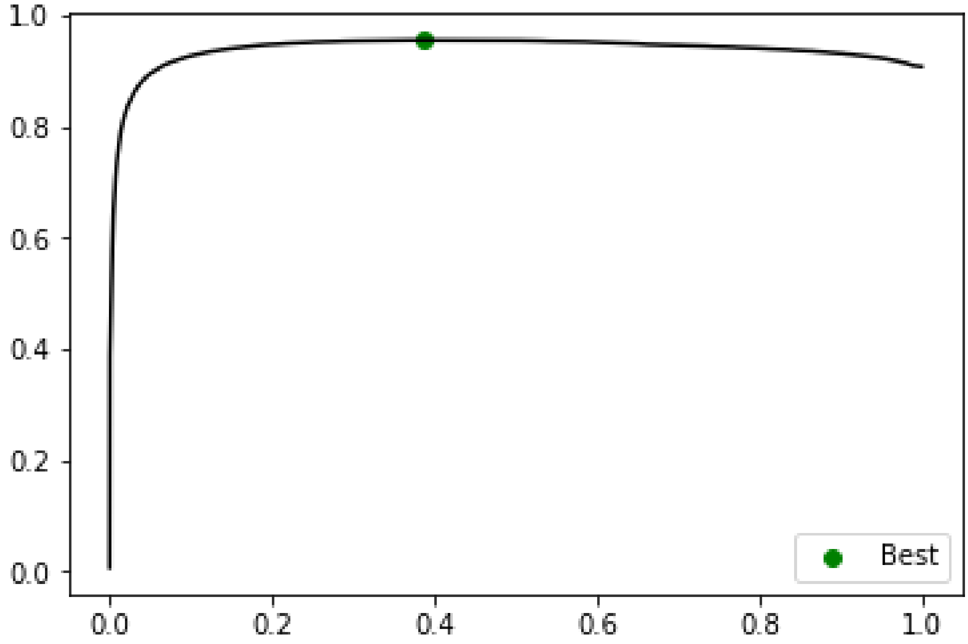
- 37_age_less_18
- 38_age_isnull
- 39_class_isnull
- 40_wage_per_hour_isnull
- 41_occupation_code_main_isnull
- 42_gains_isnull
- 43_losses_isnull
- 44_stocks_status_isnull
- 45_migration_code_change_in_msa_isnull
- 46_migration_prev_sunbelt_isnull
- 47_migration_code_move_within_reg_isnull
- 48_migration_code_change_in_reg_isnull
- 49_residence_1_year_ago_isnull
- 50_income_above_limit_isnull
- 51_gender_count
- 52_education_count
- 53_class_count
- 54_marital_status_count
- 55_race_count
- 56_is_hispanic_count
- 57_employment_commitment_count
- 58_industry_code_main_count
- 59_occupation_code_main_count
- 60_household_stat_count
- 61_household_summary_count

- 62_tax_status_count
- 63_citizenship_count
- 64_country_of_birth_own_count
- 65_country_of_birth_father_count
- 66_country_of_birth_mother_count
- 67_migration_code_change_in_msa_count
- 68_migration_prev_sunbelt_count
- 69_migration_code_move_within_reg_count
- 70_migration_code_change_in_reg_count
- 71_residence_1_year_ago_count
- 72_employment_stat_count
- 73_working_week_per_year_count
- 74_industry_code_count
- 75_occupation_code_count
- 76_total_employed_count
- 77_vet_benefit_count
- 78_mig_year_count
- 79_income_cat_count
- 80_age_mean
- 81_age_median
- 82_age_min
- 83_age_max
- 84_wage_per_hour_mean
- 85_wage_per_hour_median
- 86_wage_per_hour_min

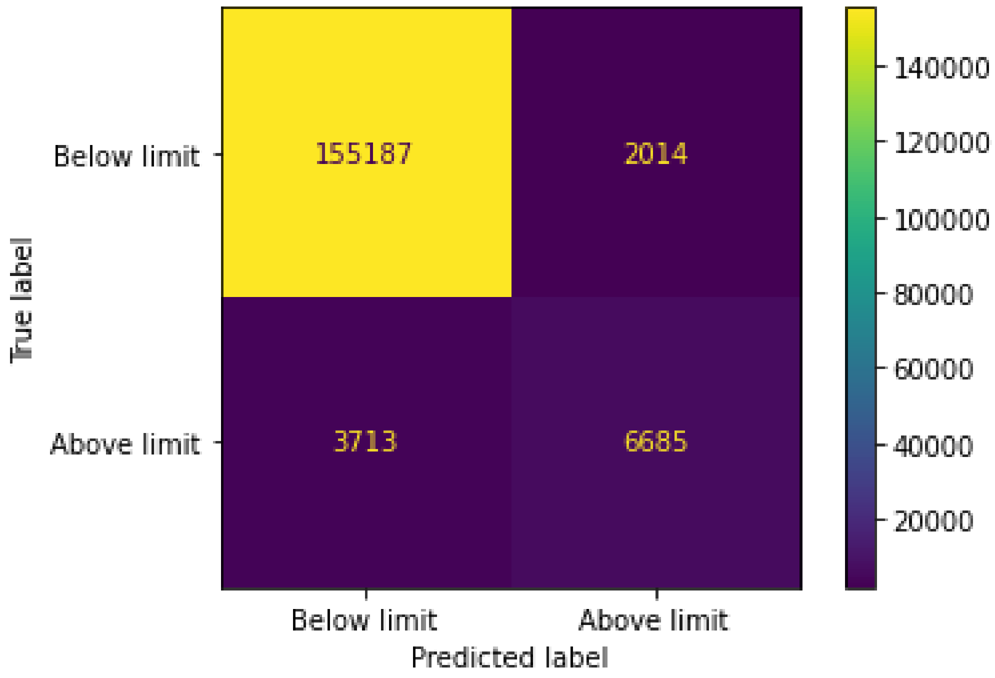
    - 87_wage_per_hour_max
    - 88_gains_mean
    - 89_gains_median
    - 90_gains_min
    - 91_gains_max
    - 92_losses_mean
    - 93_losses_median
    - 94_losses_min
    - 95_losses_max
    - 96_stocks_status_mean
    - 97_stocks_status_median
    - 98_stocks_status_min
    - 99_stocks_status_max
    - 100_importance_of_record_mean
    - 101_importance_of_record_median
    - 102_importance_of_record_min
    - 103_importance_of_record_max

- Train the catboost classifier model and evaluated with f1 metric.
- Tune the probability threshold based on the validation data.

- The optimal threshold is : 0.3862



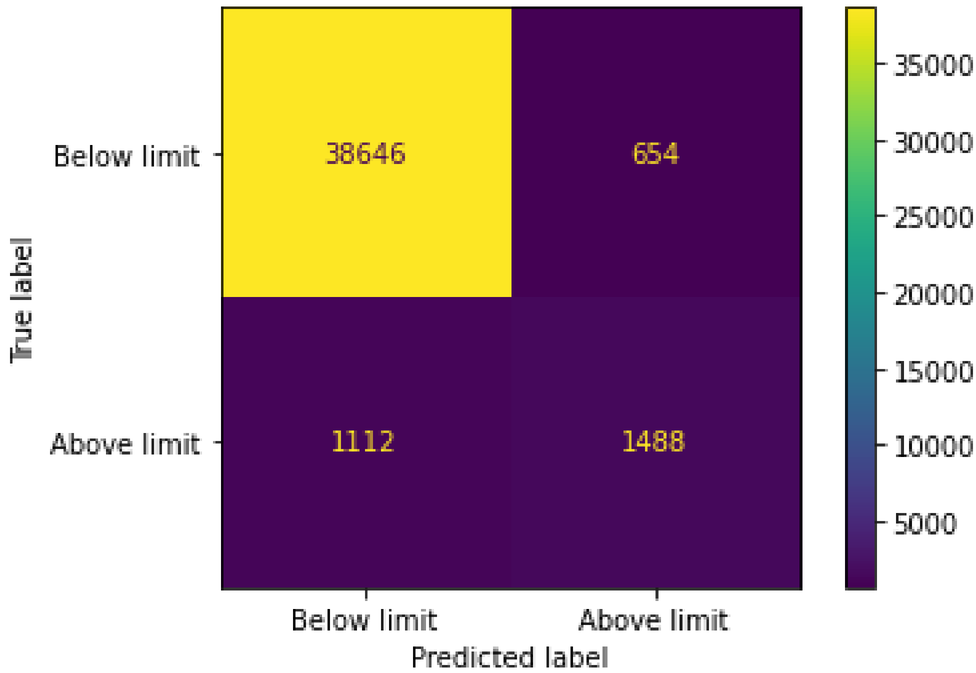
- Confusion matrix for train data



- Classification report for train data

	precision	recall	f1-score	support
Below limit	0.98	0.99	0.98	157201
Above limit	0.77	0.64	0.70	10398
accuracy			0.97	167599
macro avg	0.87	0.82	0.84	167599
weighted avg	0.96	0.97	0.96	167599

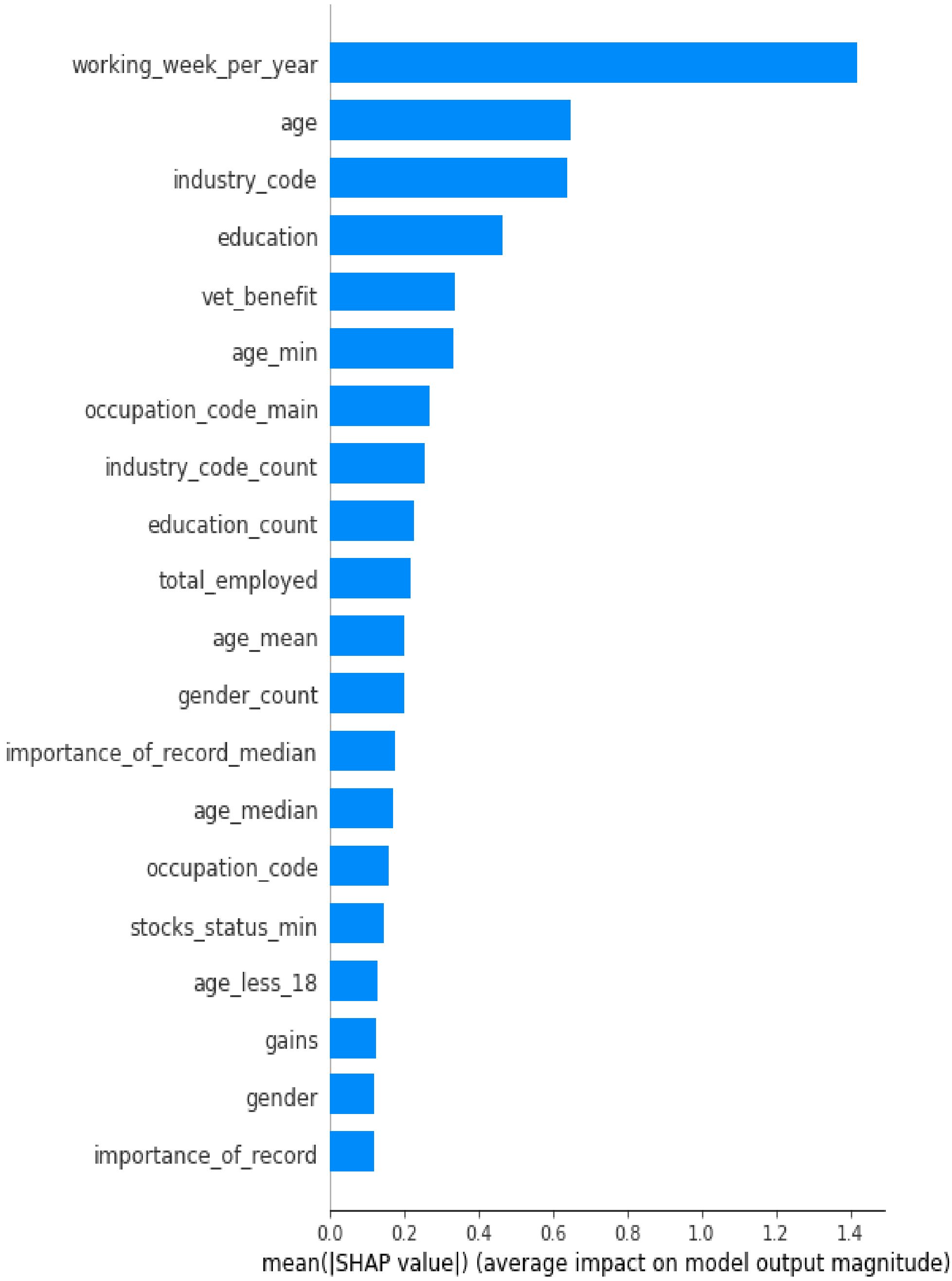
- Confusion matrix for validation data



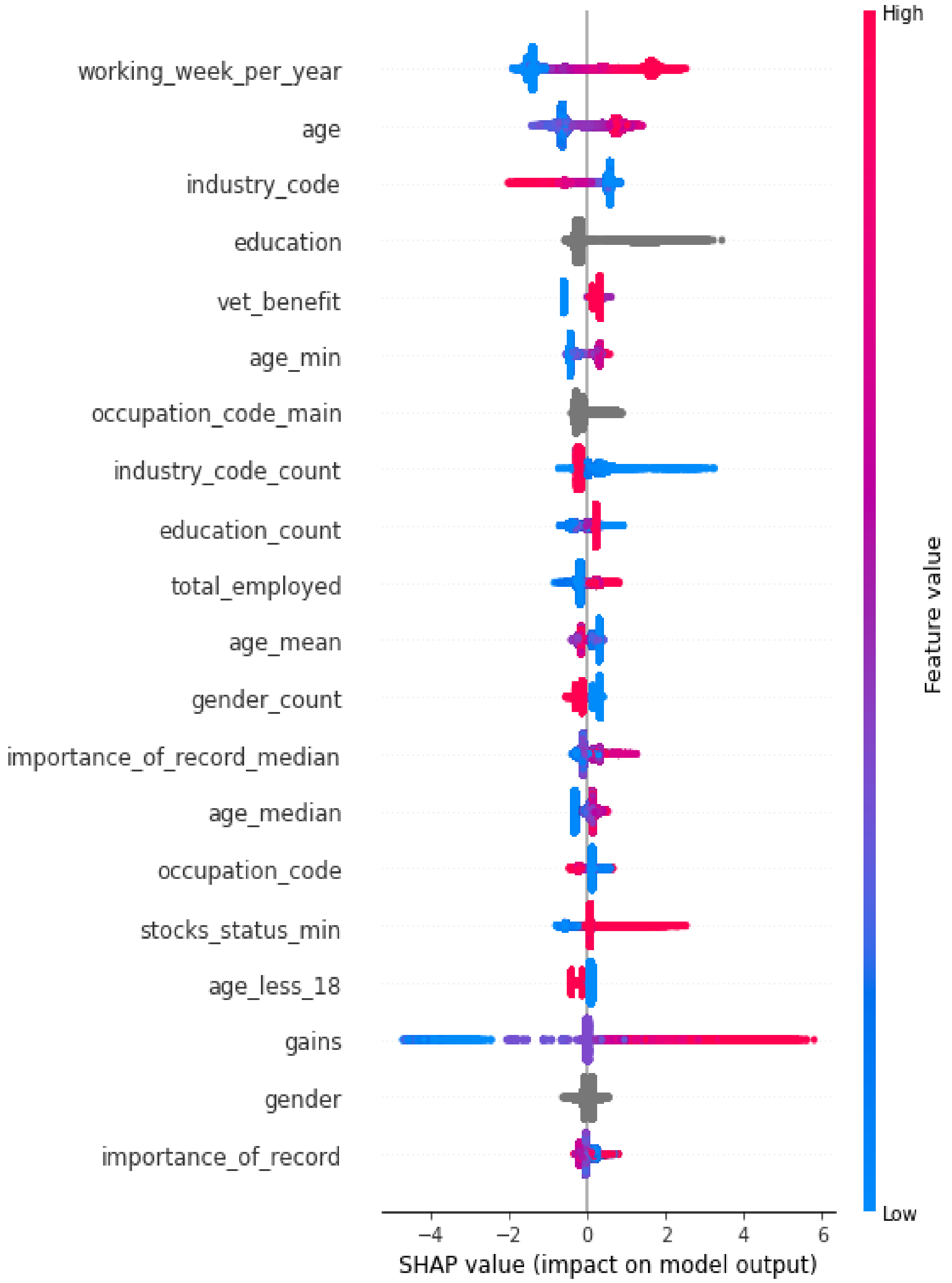
- Classification report for validation data

	precision	recall	f1-score	support
Below limit	0.97	0.98	0.98	39300
Above limit	0.69	0.57	0.63	2600
accuracy			0.96	41900
macro avg	0.83	0.78	0.80	41900
weighted avg	0.95	0.96	0.96	41900

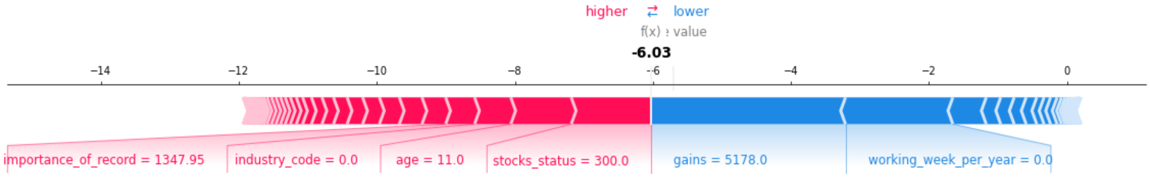
- Catboost model interpretation with SHAP
- Catboost – SHAP feature importances



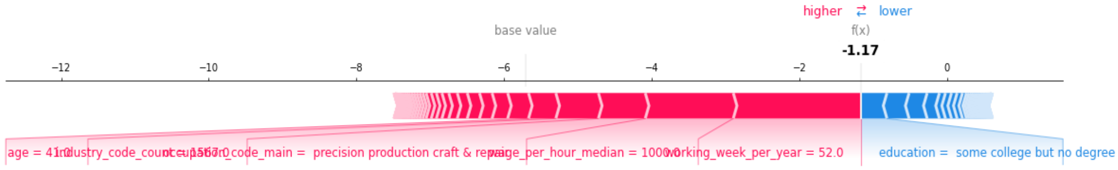
- Catboost – SHAP top feature impact



- SHAP Feature impact for single observation(class 0)



- SHAP Feature impact for single observation(class 1)



- <u>Final score is 0.613205338</u>