Approach Machinehack analytics olympiad 2022

Create a machine learning model to help an insurance company understand which claims are worth rejecting and the claims which should be accepted for reimbursement.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
  - Column name lower casr
  - Feature Engineering
    - Create a binary feature if a person having child without a marriage.
    - get a categorical columns percentage
    - numerical groupby numerical summary
- The final features for the model
  - 1_age
  - 2_gender
  - 3_driving_experience
  - 4_education
  - 5_income
  - 6_credit_score
  - 7_vehicle_ownership
  - 8_vehicle_year
  - 9_married
  - 10_children

- 11_postal_code
- 12_annual_mileage
- 13_speeding_violations
- 14_duis
- 15_past_accidents
- 16_outcome
- 17_type_of_vehicle
- 18_data
- 19_not_married_hv_child
- 20_id_age_percentage
- 21_id_gender_percentage
- 22_id_driving_experience_percentage
- 23_id_education_percentage
- 24_id_income_percentage
- 25_id_vehicle_ownership_percentage
- 26_id_vehicle_year_percentage
- 27_id_married_percentage
- 28_id_children_percentage
- 29_id_type_of_vehicle_percentage
- 30_id_not_married_hv_child_percentage
- 31_id_annual_mileage_percentage
- 32_id_speeding_violations_percentage
- 33_id_duis_percentage
- 34_id_past_accidents_percentage
- 35_age_credit_score_median

- 36_age_credit_score_mean
- 37_age_credit_score_max
- 38_age_credit_score_min
- 39_age_annual_mileage_median
- 40_age_annual_mileage_mean
- 41_age_annual_mileage_max
- 42_age_annual_mileage_min
- 43_age_speeding_violations_median
- 44_age_speeding_violations_mean
- 45_age_speeding_violations_max
- 46_age_speeding_violations_min
- 47_age_duis_median
- 48_age_duis_mean
- 49_age_duis_max
- 50_age_duis_min
- 51_age_past_accidents_median
- 52_age_past_accidents_mean
- 53_age_past_accidents_max
- 54_age_past_accidents_min
- 55_gender_credit_score_median
- 56_gender_credit_score_mean
- 57_gender_credit_score_max
- 58_gender_credit_score_min
- 59_gender_annual_mileage_median
- 60_gender_annual_mileage_mean

- 61_gender_annual_mileage_max
- 62_gender_annual_mileage_min
- 63_gender_speeding_violations_median
- 64_gender_speeding_violations_mean
- 65_gender_speeding_violations_max
- 66_gender_speeding_violations_min
- 67_gender_duis_median
- 68_gender_duis_mean
- 69_gender_duis_max
- 70_gender_duis_min
- 71_gender_past_accidents_median
- 72_gender_past_accidents_mean
- 73_gender_past_accidents_max
- 74_gender_past_accidents_min
- 75_driving_experience_credit_score_median
- 76_driving_experience_credit_score_mean
- 77_driving_experience_credit_score_max
- 78_driving_experience_credit_score_min
- 79_driving_experience_annual_mileage_median
- 80_driving_experience_annual_mileage_mean
- 81_driving_experience_annual_mileage_max
- 82_driving_experience_annual_mileage_min
- 83_driving_experience_speeding_violations_median
- 84_driving_experience_speeding_violations_mean
- 85_driving_experience_speeding_violations_max

- 86_driving_experience_speeding_violations_min
- 87_driving_experience_duis_median
- 88_driving_experience_duis_mean
- 89_driving_experience_duis_max
- 90_driving_experience_duis_min
- 91_driving_experience_past_accidents_median
- 92_driving_experience_past_accidents_mean
- 93_driving_experience_past_accidents_max
- 94_driving_experience_past_accidents_min
- 95_education_credit_score_median
- 96_education_credit_score_mean
- 97_education_credit_score_max
- 98_education_credit_score_min
- 99_education_annual_mileage_median
- 100_education_annual_mileage_mean
- 101_education_annual_mileage_max
- 102_education_annual_mileage_min
- 103_education_speeding_violations_median
- 104_education_speeding_violations_mean
- 105_education_speeding_violations_max
- 106_education_speeding_violations_min
- 107_education_duis_median
- 108_education_duis_mean
- 109_education_duis_max
- 110_education_duis_min

- 111_education_past_accidents_median
- 112_education_past_accidents_mean
- 113_education_past_accidents_max
- 114_education_past_accidents_min
- 115_income_credit_score_median
- 116_income_credit_score_mean
- 117_income_credit_score_max
- 118_income_credit_score_min
- 119_income_annual_mileage_median
- 120_income_annual_mileage_mean
- 121_income_annual_mileage_max
- 122_income_annual_mileage_min
- 123_income_speeding_violations_median
- 124_income_speeding_violations_mean
- 125_income_speeding_violations_max
- 126_income_speeding_violations_min
- 127_income_duis_median
- 128_income_duis_mean
- 129_income_duis_max
- 130_income_duis_min
- 131_income_past_accidents_median
- 132_income_past_accidents_mean
- 133_income_past_accidents_max
- 134_income_past_accidents_min
- 135_vehicle_ownership_credit_score_median

- 136_vehicle_ownership_credit_score_mean
- 137_vehicle_ownership_credit_score_max
- 138_vehicle_ownership_credit_score_min
- 139_vehicle_ownership_annual_mileage_median
- 140_vehicle_ownership_annual_mileage_mean
- 141_vehicle_ownership_annual_mileage_max
- 142_vehicle_ownership_annual_mileage_min
- 143_vehicle_ownership_speeding_violations_median
- 144_vehicle_ownership_speeding_violations_mean
- 145_vehicle_ownership_speeding_violations_max
- 146_vehicle_ownership_speeding_violations_min
- 147_vehicle_ownership_duis_median
- 148_vehicle_ownership_duis_mean
- 149_vehicle_ownership_duis_max
- 150_vehicle_ownership_duis_min
- 151_vehicle_ownership_past_accidents_median
- 152_vehicle_ownership_past_accidents_mean
- 153_vehicle_ownership_past_accidents_max
- 154_vehicle_ownership_past_accidents_min
- 155_vehicle_year_credit_score_median
- 156_vehicle_year_credit_score_mean
- 157_vehicle_year_credit_score_max
- 158_vehicle_year_credit_score_min
- 159_vehicle_year_annual_mileage_median
- 160_vehicle_year_annual_mileage_mean

- 161_vehicle_year_annual_mileage_max
- 162_vehicle_year_annual_mileage_min
- 163_vehicle_year_speeding_violations_median
- 164_vehicle_year_speeding_violations_mean
- 165_vehicle_year_speeding_violations_max
- 166_vehicle_year_speeding_violations_min
- 167_vehicle_year_duis_median
- 168_vehicle_year_duis_mean
- 169_vehicle_year_duis_max
- 170_vehicle_year_duis_min
- 171_vehicle_year_past_accidents_median
- 172_vehicle_year_past_accidents_mean
- 173_vehicle_year_past_accidents_max
- 174_vehicle_year_past_accidents_min
- 175_married_credit_score_median
- 176_married_credit_score_mean
- 177_married_credit_score_max
- 178_married_credit_score_min
- 179_married_annual_mileage_median
- 180_married_annual_mileage_mean
- 181_married_annual_mileage_max
- 182_married_annual_mileage_min
- 183_married_speeding_violations_median
- 184_married_speeding_violations_mean
- 185_married_speeding_violations_max

- 186_married_speeding_violations_min
- 187_married_duis_median
- 188_married_duis_mean
- 189_married_duis_max
- 190_married_duis_min
- 191_married_past_accidents_median
- 192_married_past_accidents_mean
- 193_married_past_accidents_max
- 194_married_past_accidents_min
- 195_not_married_hv_child_credit_score_median
- 196_not_married_hv_child_credit_score_mean
- 197_not_married_hv_child_credit_score_max
- 198_not_married_hv_child_credit_score_min
- 199_not_married_hv_child_annual_mileage_median
- 200_not_married_hv_child_annual_mileage_mean
- 201_not_married_hv_child_annual_mileage_max
- 202_not_married_hv_child_annual_mileage_min
- 203_not_married_hv_child_speeding_violations_median
- 204_not_married_hv_child_speeding_violations_mean
- 205_not_married_hv_child_speeding_violations_max
- 206_not_married_hv_child_speeding_violations_min
- 207_not_married_hv_child_duis_median
- 208_not_married_hv_child_duis_mean
- 209_not_married_hv_child_duis_max
- 210_not_married_hv_child_duis_min

- 211_not_married_hv_child_past_accidents_median
- 212_not_married_hv_child_past_accidents_mean
- 213_not_married_hv_child_past_accidents_max
- 214_not_married_hv_child_past_accidents_min
- 215_children_credit_score_median
- 216_children_credit_score_mean
- 217_children_credit_score_max
- 218_children_credit_score_min
- 219_children_annual_mileage_median
- 220_children_annual_mileage_mean
- 221_children_annual_mileage_max
- 222_children_annual_mileage_min
- 223_children_speeding_violations_median
- 224_children_speeding_violations_mean
- 225_children_speeding_violations_max
- 226_children_speeding_violations_min
- 227_children_duis_median
- 228_children_duis_mean
- 229_children_duis_max
- 230_children_duis_min
- 231_children_past_accidents_median
- 232_children_past_accidents_mean
- 233_children_past_accidents_max
- 234_children_past_accidents_min
- 235_type_of_vehicle_credit_score_median

- 236_type_of_vehicle_credit_score_mean
- 237_type_of_vehicle_credit_score_max
- 238_type_of_vehicle_credit_score_min
- 239_type_of_vehicle_annual_mileage_median
- 240_type_of_vehicle_annual_mileage_mean
- 241_type_of_vehicle_annual_mileage_max
- 242_type_of_vehicle_annual_mileage_min
- 243_type_of_vehicle_speeding_violations_median
- 244_type_of_vehicle_speeding_violations_mean
- 245_type_of_vehicle_speeding_violations_max
- 246_type_of_vehicle_speeding_violations_min
- 247_type_of_vehicle_duis_median
- 248_type_of_vehicle_duis_mean
- 249_type_of_vehicle_duis_max
- 250_type_of_vehicle_duis_min
- 251_type_of_vehicle_past_accidents_median
- 252_type_of_vehicle_past_accidents_mean
- 253_type_of_vehicle_past_accidents_max
- 254_type_of_vehicle_past_accidents_min

- Created catboost classifier model and tuned hyperparameters by using optuna framework.Model evaluated with Logloss.After 100 trials,
    - The best score is 0.68106
    - The best hyperparemeters are,

      {'reg_lambda': 160,

      'learning_rate': 0.04010653349368823,

      'n_estimators': 221,

      'max_depth': 6,

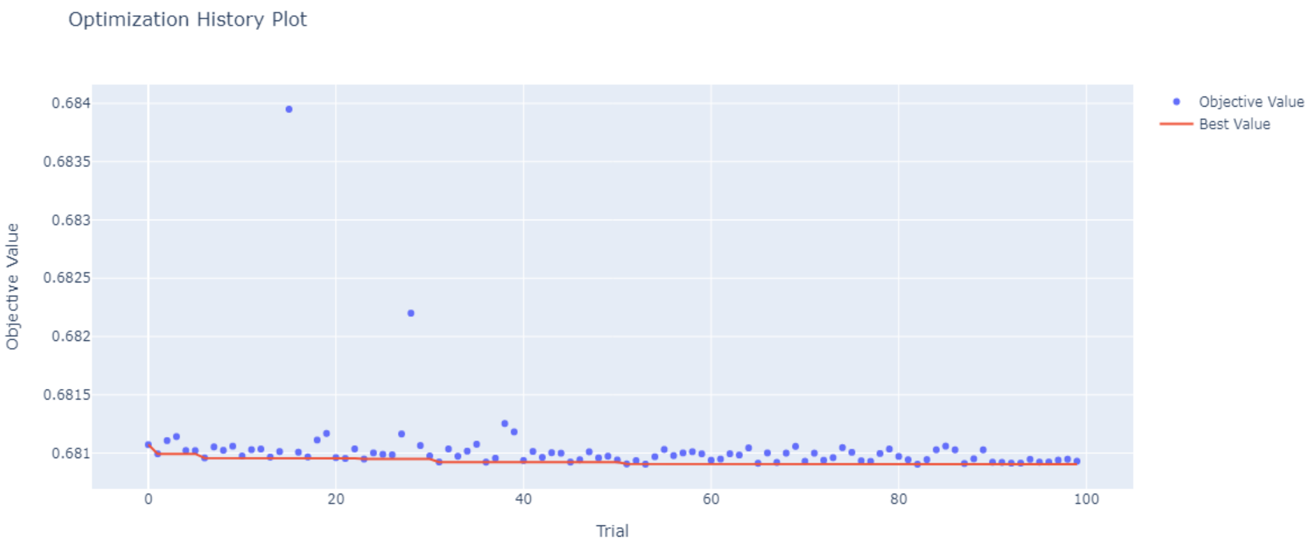
      'od_type': 'IncToDec',

      'random_state': 1024,

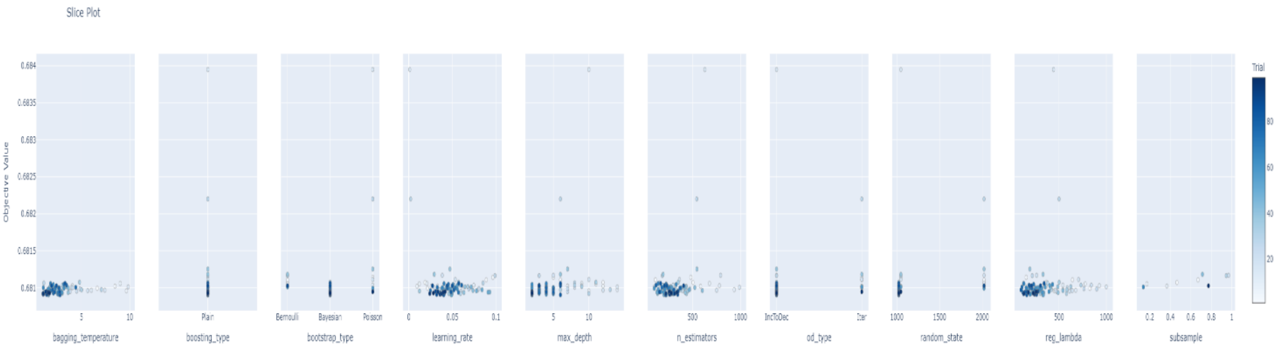
      'boosting_type': 'Plain',

      'bootstrap_type': 'Bayesian',

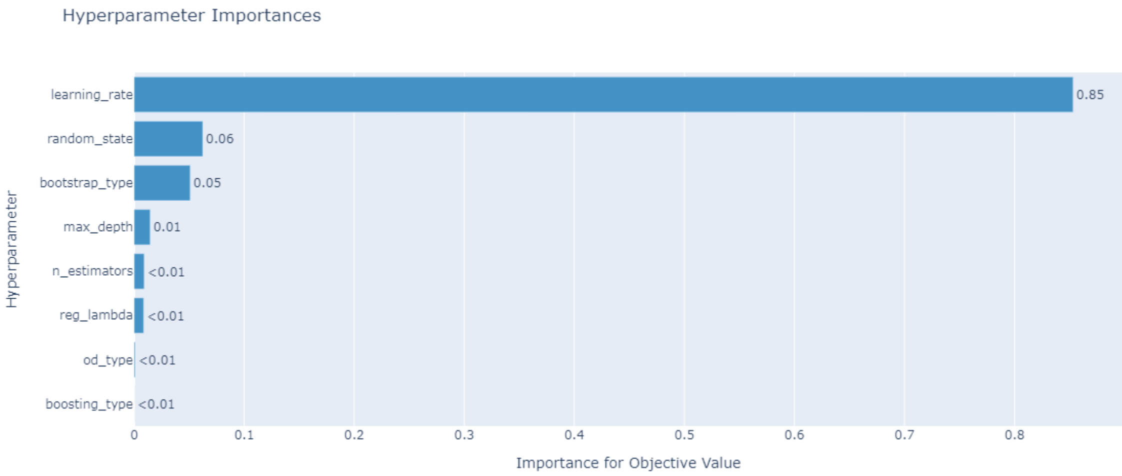
      'bagging_temperature': 2.7779687356027973}
- Visualizing the Optimization History - Explains the best score at each trials.



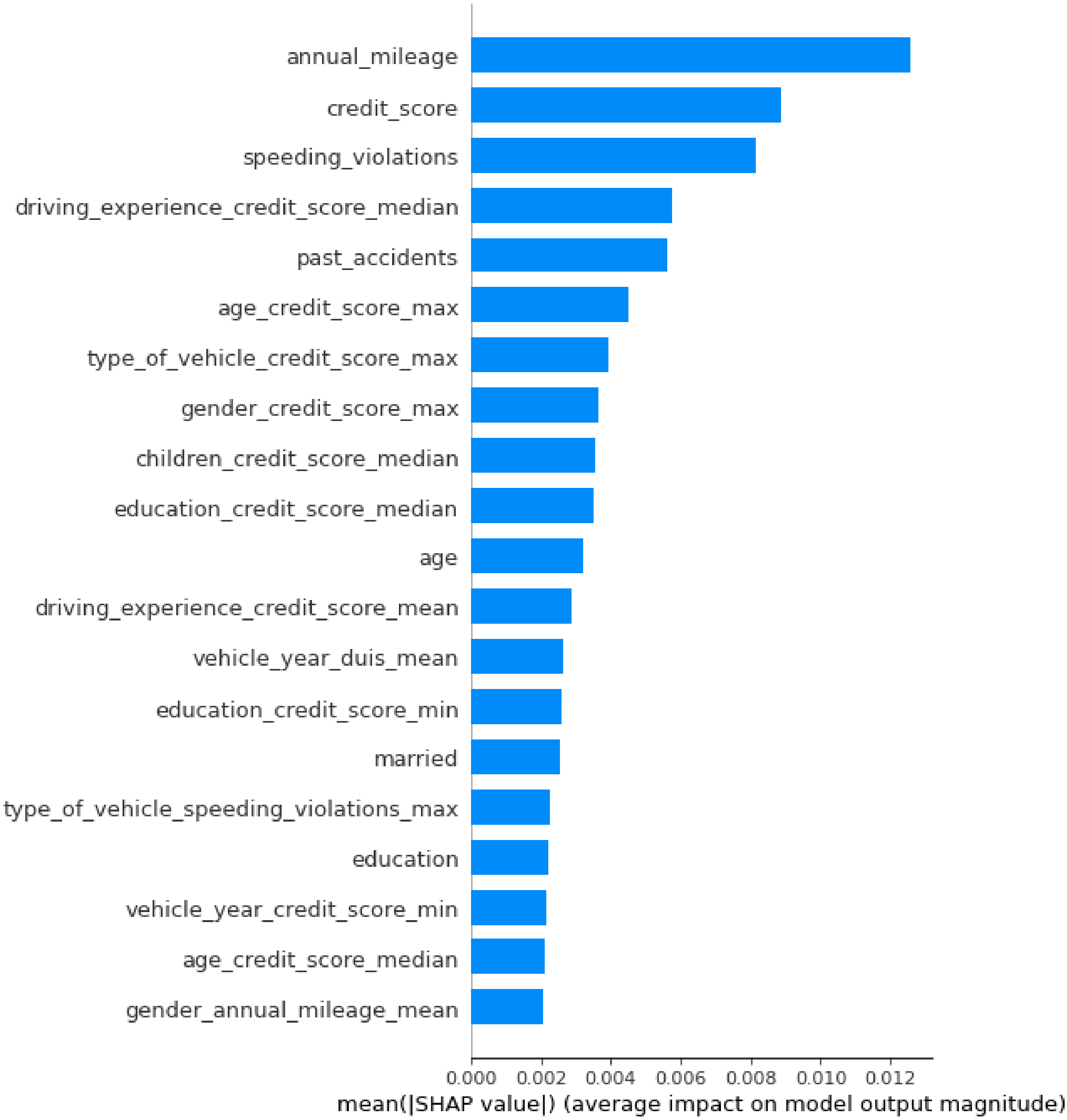
- Visualizing High-dimensional Parameter Relationships



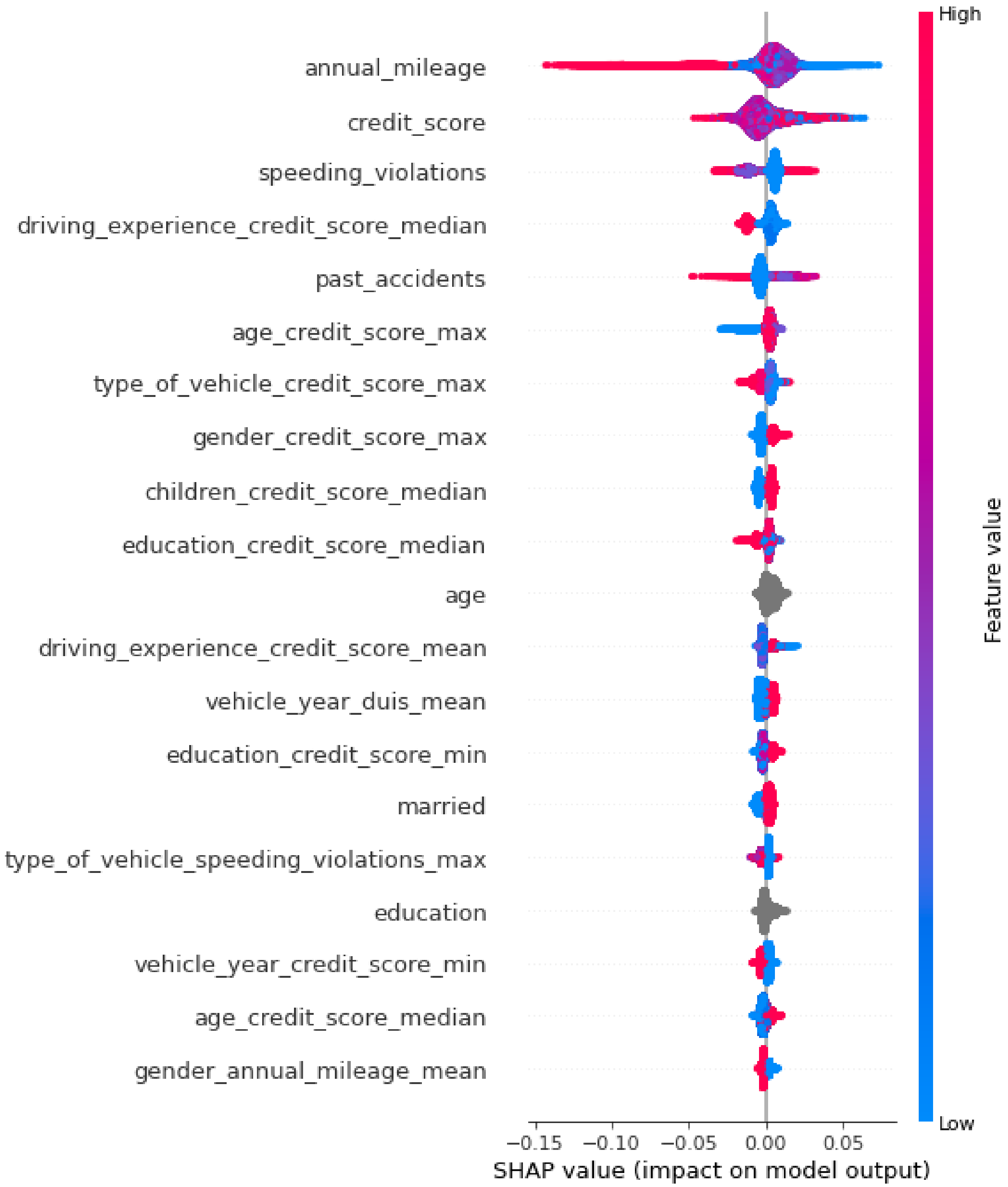
- Visualizing Parameter Importances



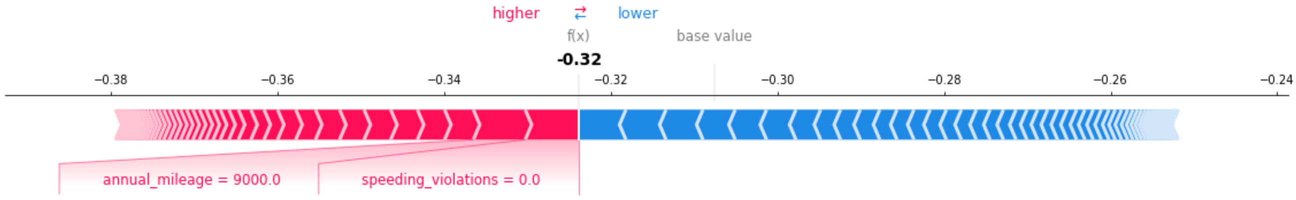
- Catboost SHAP feature importances plot



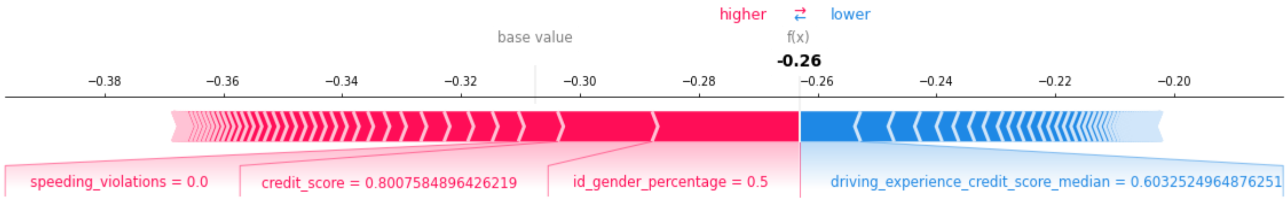
- Catboost SHAP top features impact the model



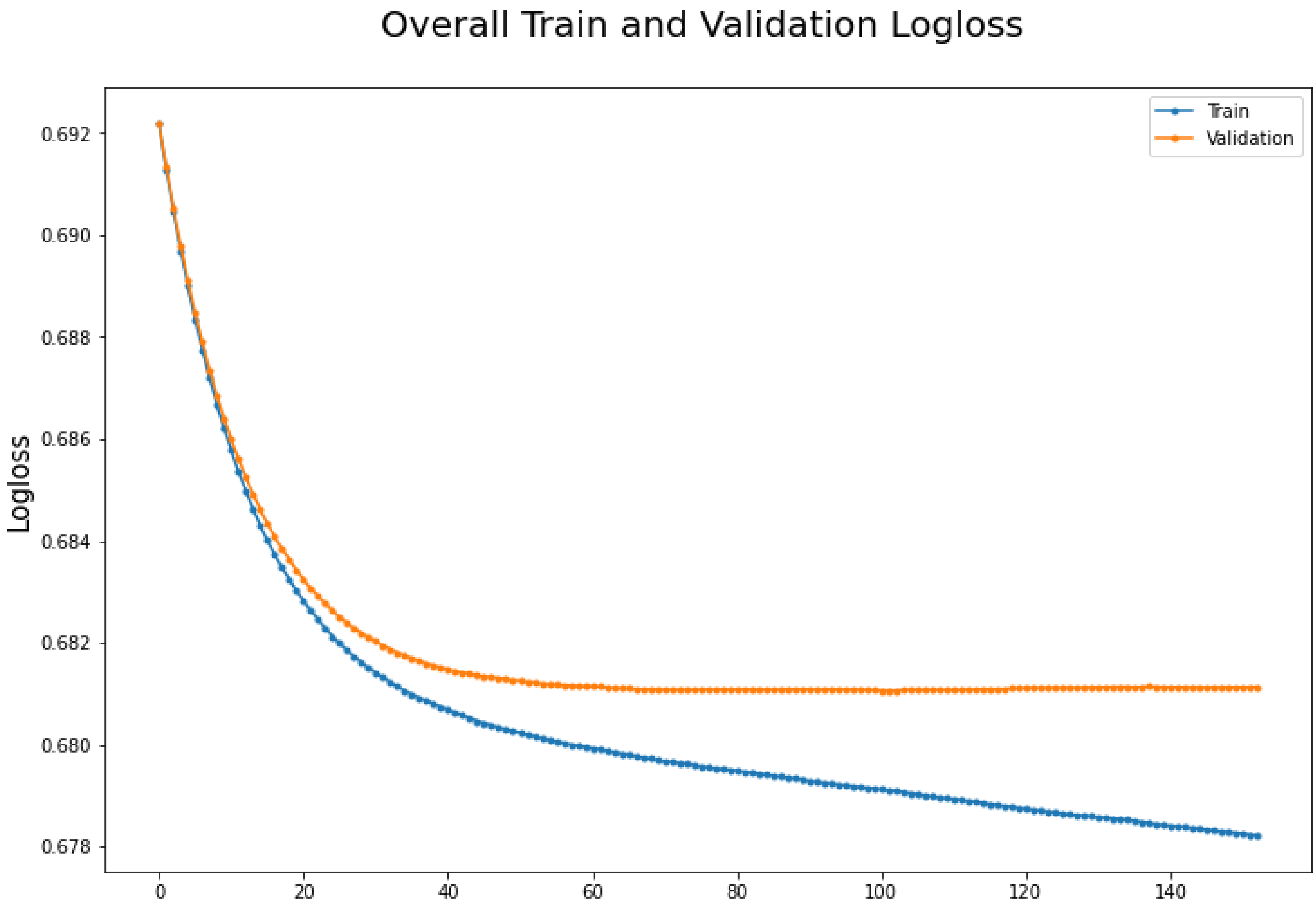
- Top feature influences for class 1



- Top feature influences for class 0



- Overall Train and Validation Logloss



- Train logloss: 0.67936 , Validation logloss: 0.68106
- Final competition score is 0.68081