Approach_AV Dataverse Hack - Insurance Claim Prediction

Create a machine learning model to predict if the policyholder will file a claim in the next 6 months or not based on the set of car and policy features.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
    - Feature Engineering
        - From max_torque split the newton meter and rpm value then convert them into numeric.
        - From max_power split the bhp and rpm value then convert them into numeric.
        - Convert the engine displacement cc to litre.
        - Create car brand name feature from engine type
        - Map the 1 and 0 for the yes/no value categorical columns.
        - Based on the area cluster create numerical summary for the population density column.
        - Get groupby numerical summary for the columns,
            - Policy_tenure
            - Age_of_car
            - Age_of_policyholder
        - Z-score outlier detection for numerical columns.

- The final features for the model
    - 0_policy_tenure
    - 1_age_of_car
    - 2_age_of_policyholder
    - 3_area_cluster
    - 4_population_density
    - 5_make
    - 6_segment
    - 7_model
    - 8_fuel_type
    - 9_engine_type
    - 10_airbags
    - 11_is_esc
    - 12_is_adjustable_steering
    - 13_is_tpms
    - 14_is_parking_sensors
    - 15_is_parking_camera
    - 16_rear_brakes_type
    - 17_cylinder
    - 18_transmission_type
    - 19_gear_box
    - 20_steering_type
    - 21_turning_radius
    - 22_length
    - 23_width

24_height

25_gross_weight

26_is_front_fog_lights

27_is_rear_window_wiper

28_is_rear_window_washer

29_is_rear_window_defogger

30_is_brake_assist

31_is_power_door_locks

32_is_central_locking

33_is_power_steering

34_is_driver_seat_height_adjustable

35_is_day_night_rear_view_mirror

36_is_ecw

37_is_speed_alert

38_ncap_rating

39_max_torque_nm

40_max_torque_rpm

41_max_power_bhp

42_max_power_rpm

43_engine

44_car_brand

45_population_density_min

46_population_density_max

47_population_density_mean

48_population_density_median

49_make_policy_tenure_median
50_make_policy_tenure_max
51_make_policy_tenure_min
52_make_policy_tenure_mean
53_make_age_of_car_median
54_make_age_of_car_max
55_make_age_of_car_min
56_make_age_of_car_mean
57_make_age_of_policyholder_median
58_make_age_of_policyholder_max
59_make_age_of_policyholder_min
60_make_age_of_policyholder_mean
61_segment_policy_tenure_median
62_segment_policy_tenure_max
63_segment_policy_tenure_min
64_segment_policy_tenure_mean
65_segment_age_of_car_median
66_segment_age_of_car_max
67_segment_age_of_car_min
68_segment_age_of_car_mean
69_segment_age_of_policyholder_median
70_segment_age_of_policyholder_max
71_segment_age_of_policyholder_min
72_segment_age_of_policyholder_mean
73_model_policy_tenure_median

74_model_policy_tenure_max
75_model_policy_tenure_min
76_model_policy_tenure_mean
77_model_age_of_car_median
78_model_age_of_car_max
79_model_age_of_car_min
80_model_age_of_car_mean
81_model_age_of_policyholder_median
82_model_age_of_policyholder_max
83_model_age_of_policyholder_min
84_model_age_of_policyholder_mean
85_fuel_type_policy_tenure_median
86_fuel_type_policy_tenure_max
87_fuel_type_policy_tenure_min
88_fuel_type_policy_tenure_mean
89_fuel_type_age_of_car_median
90_fuel_type_age_of_car_max
91_fuel_type_age_of_car_min
92_fuel_type_age_of_car_mean
93_fuel_type_age_of_policyholder_median
94_fuel_type_age_of_policyholder_max
95_fuel_type_age_of_policyholder_min
96_fuel_type_age_of_policyholder_mean
97_rear_brakes_type_policy_tenure_median
98_rear_brakes_type_policy_tenure_max

99_rear_brakes_type_policy_tenure_min
100_rear_brakes_type_policy_tenure_mean
101_rear_brakes_type_age_of_car_median
102_rear_brakes_type_age_of_car_max
103_rear_brakes_type_age_of_car_min
104_rear_brakes_type_age_of_car_mean
105_rear_brakes_type_age_of_policyholder_median
106_rear_brakes_type_age_of_policyholder_max
107_rear_brakes_type_age_of_policyholder_min
108_rear_brakes_type_age_of_policyholder_mean
109_engine_policy_tenure_median
110_engine_policy_tenure_max
111_engine_policy_tenure_min
112_engine_policy_tenure_mean
113_engine_age_of_car_median
114_engine_age_of_car_max
115_engine_age_of_car_min
116_engine_age_of_car_mean
117_engine_age_of_policyholder_median
118_engine_age_of_policyholder_max
119_engine_age_of_policyholder_min
120_engine_age_of_policyholder_mean
121_cylinder_policy_tenure_median
122_cylinder_policy_tenure_max
123_cylinder_policy_tenure_min

124_cylinder_policy_tenure_mean
125_cylinder_age_of_car_median
126_cylinder_age_of_car_max
127_cylinder_age_of_car_min
128_cylinder_age_of_car_mean
129_cylinder_age_of_policyholder_median
130_cylinder_age_of_policyholder_max
131_cylinder_age_of_policyholder_min
132_cylinder_age_of_policyholder_mean
133_transmission_type_policy_tenure_median
134_transmission_type_policy_tenure_max
135_transmission_type_policy_tenure_min
136_transmission_type_policy_tenure_mean
137_transmission_type_age_of_car_median
138_transmission_type_age_of_car_max
139_transmission_type_age_of_car_min
140_transmission_type_age_of_car_mean
141_transmission_type_age_of_policyholder_median
142_transmission_type_age_of_policyholder_max
143_transmission_type_age_of_policyholder_min
144_transmission_type_age_of_policyholder_mean
145_gear_box_policy_tenure_median
146_gear_box_policy_tenure_max
147_gear_box_policy_tenure_min
148_gear_box_policy_tenure_mean

149_gear_box_age_of_car_median
150_gear_box_age_of_car_max
151_gear_box_age_of_car_min
152_gear_box_age_of_car_mean
153_gear_box_age_of_policyholder_median
154_gear_box_age_of_policyholder_max
155_gear_box_age_of_policyholder_min
156_gear_box_age_of_policyholder_mean
157_steering_type_policy_tenure_median
158_steering_type_policy_tenure_max
159_steering_type_policy_tenure_min
160_steering_type_policy_tenure_mean
161_steering_type_age_of_car_median
162_steering_type_age_of_car_max
163_steering_type_age_of_car_min
164_steering_type_age_of_car_mean
165_steering_type_age_of_policyholder_median
166_steering_type_age_of_policyholder_max
167_steering_type_age_of_policyholder_min
168_steering_type_age_of_policyholder_mean
169_car_brand_policy_tenure_median
170_car_brand_policy_tenure_max
171_car_brand_policy_tenure_min
172_car_brand_policy_tenure_mean
173_car_brand_age_of_car_median

174_car_brand_age_of_car_max

175_car_brand_age_of_car_min

176_car_brand_age_of_car_mean

177_car_brand_age_of_policyholder_median

178_car_brand_age_of_policyholder_max

179_car_brand_age_of_policyholder_min

180_car_brand_age_of_policyholder_mean

181_policy_tenure_outlier

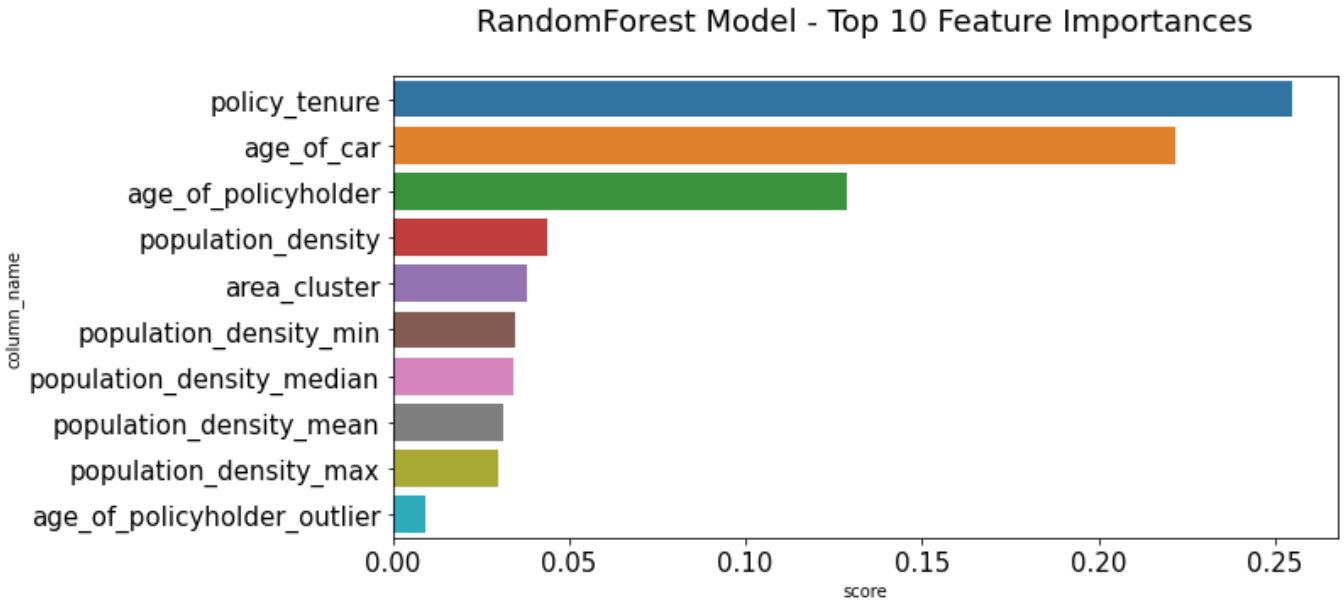
182_age_of_car_outlier

183_age_of_policyholder_outlier

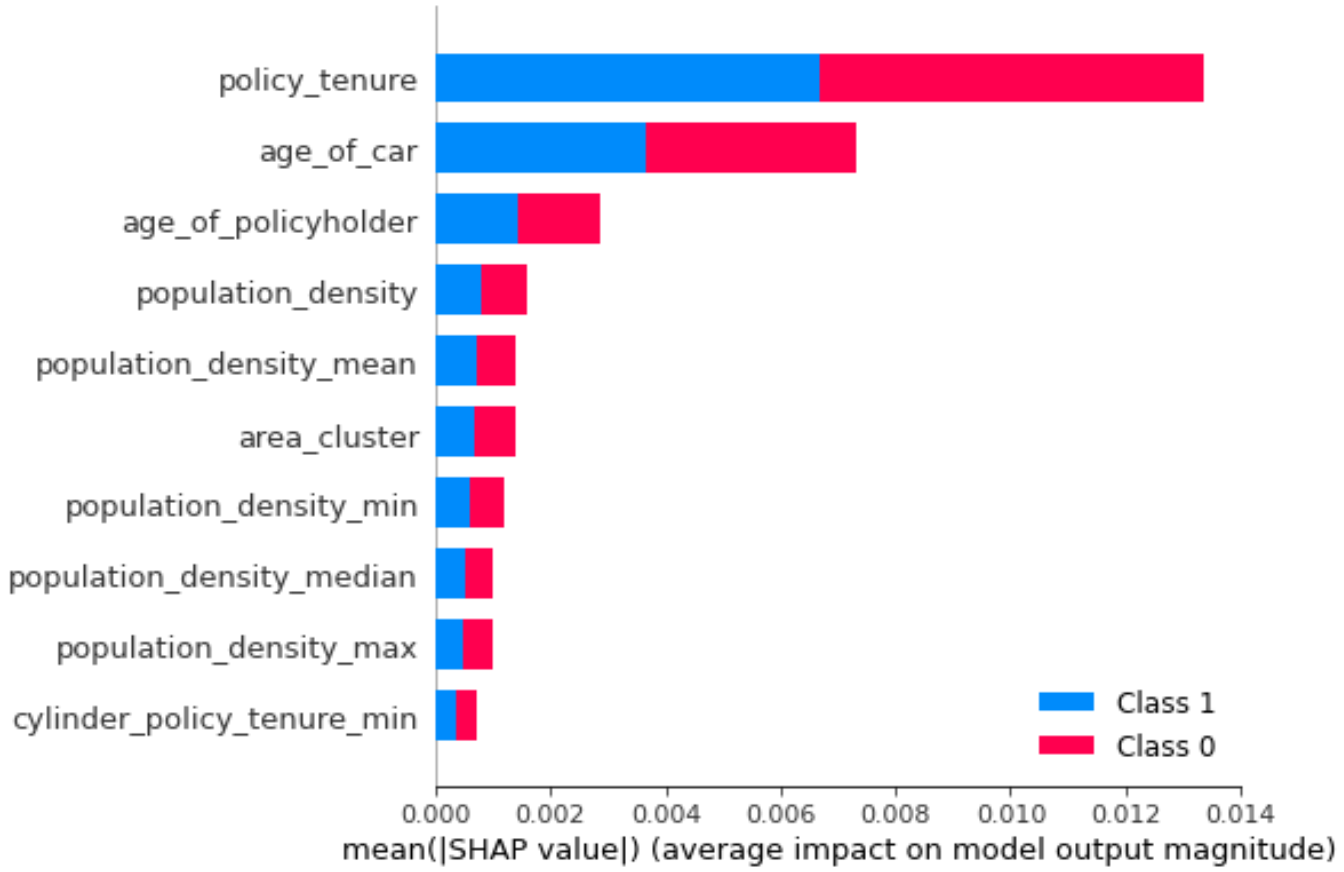
- Created Random Forest classifier model with maximum depth 6.Model evaluated with f1-score.,
    - The model paremeters are,
        - bootstrap:True
        - ccp_alpha:0.0
        - class_weight:None
        - criterion:gini
        - max_depth:6
        - max_features:auto
        - max_leaf_nodes:None
        - max_samples:None
        - min_impurity_decrease:0.0
        - min_samples_leaf:1
        - min_samples_split:2

- min_weight_fraction_leaf:0.0
- n_estimators:100
- n_jobs:None
- oob_score:False
- random_state:42
- verbose:0
- warm_start:False

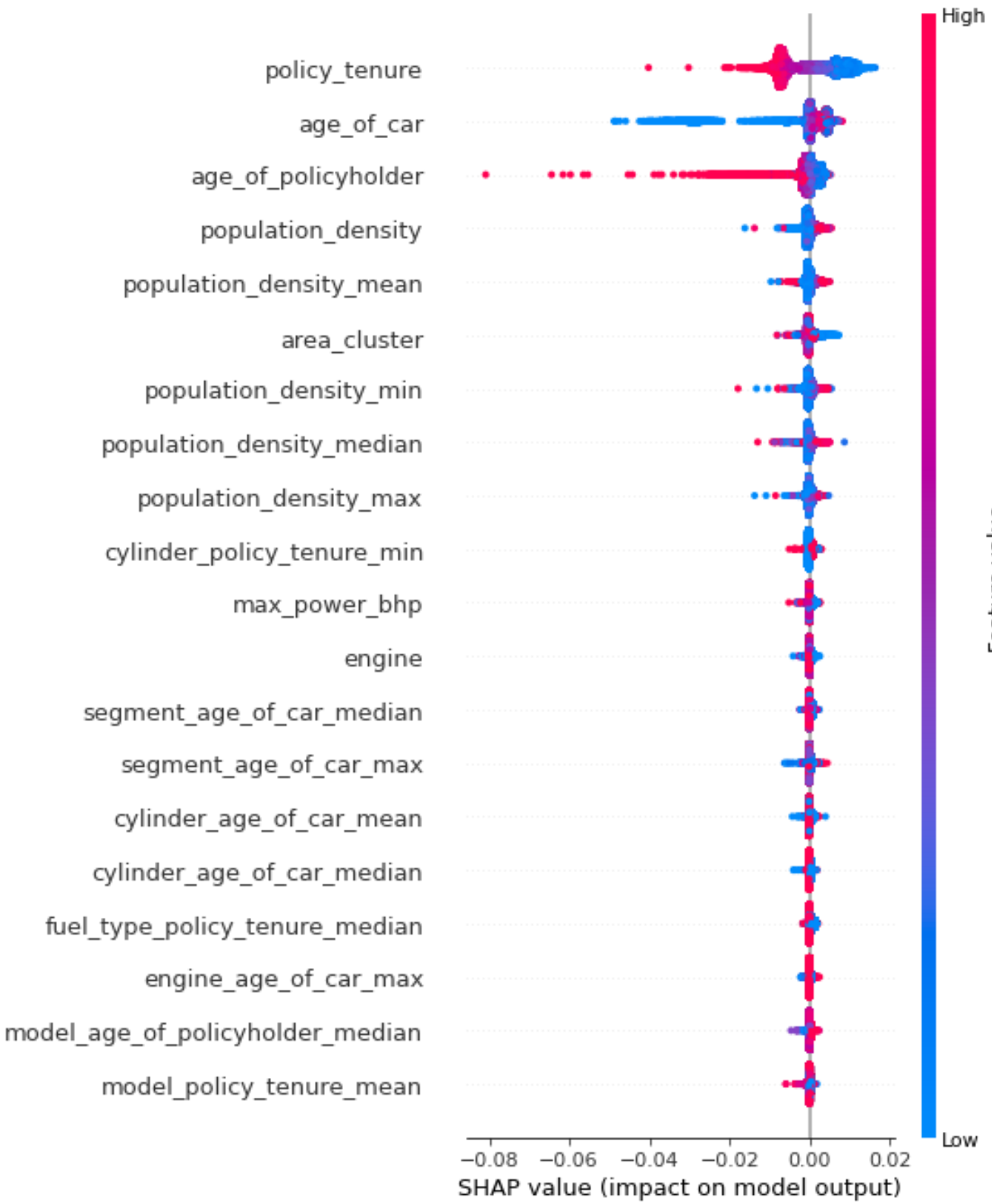
- RandomForest Model - Top 10 Feature Importances



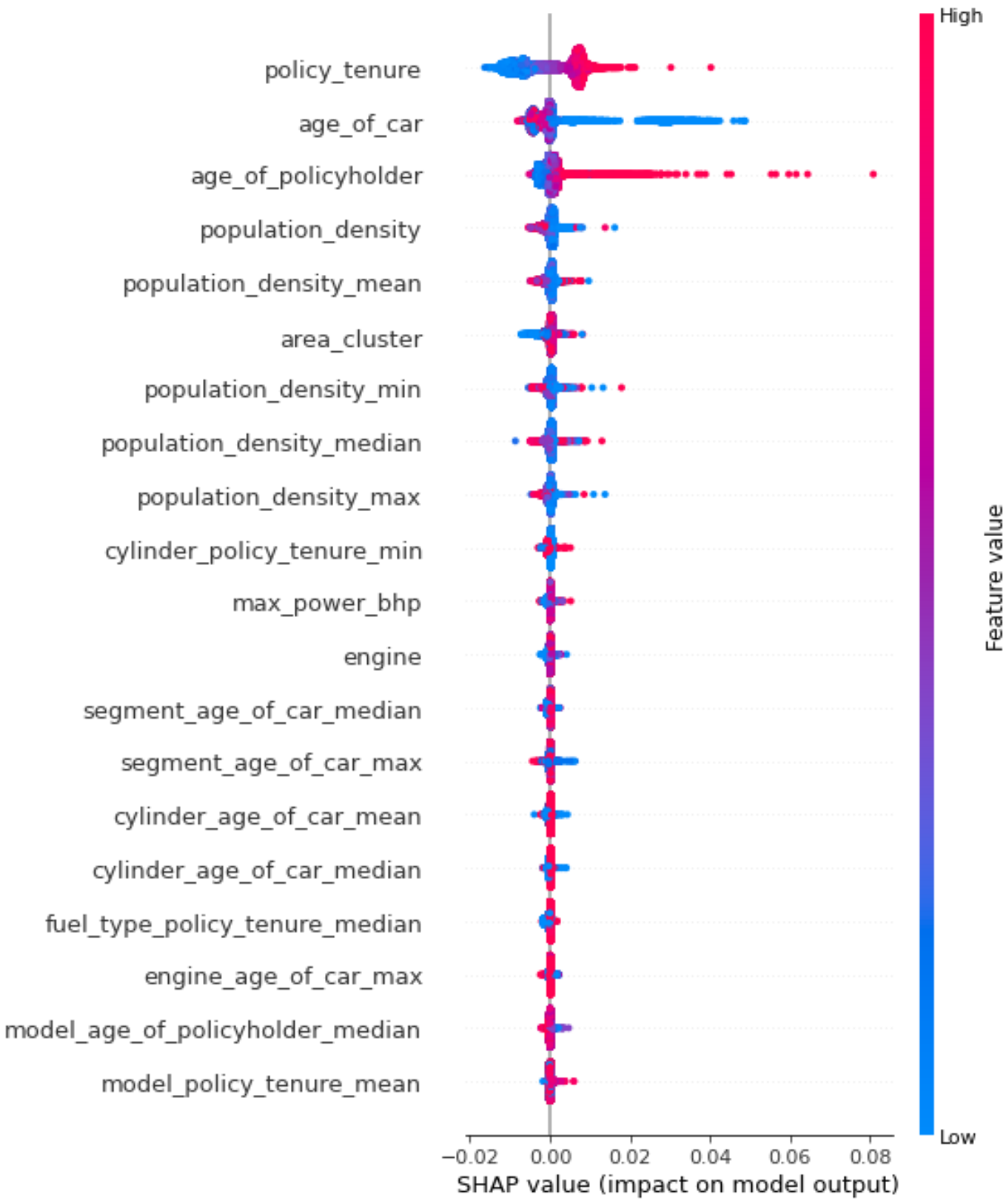
- Randomforest model - SHAP top 10 feature importances



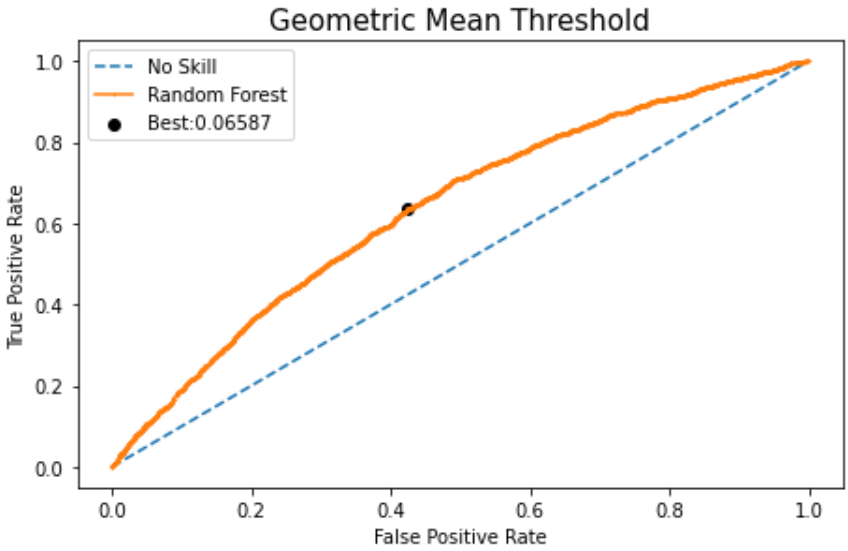
- SHAP Top feature influences the class 0



- SHAP Top feature influences the class 1



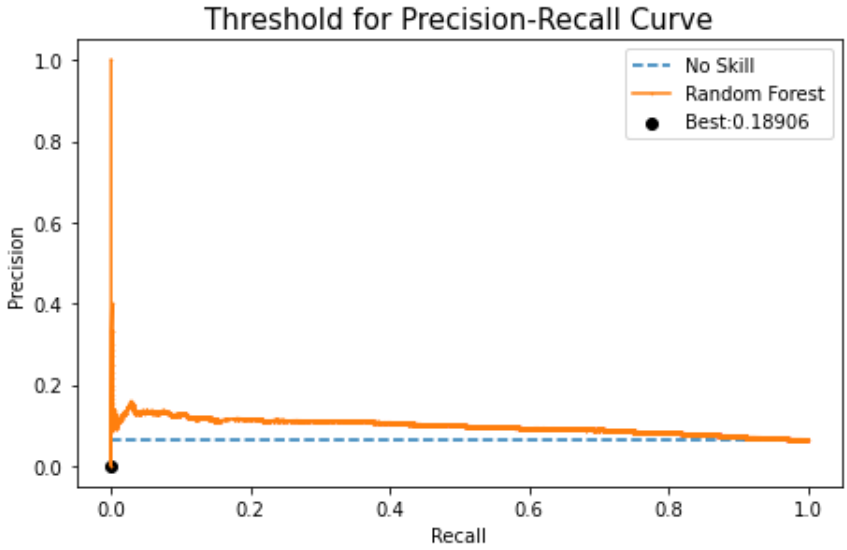
- Threshold Tuning Results
    - Geometric Mean Threshold - Best Threshold 0.06587, G-Mean 0.60513



Confusion Matrix

	0	1	accuracy	macro avg	weighted avg
precision	0.9585	0.09282	0.57939	0.52566	0.90312
recall	0.57556	0.63541	0.57939	0.60548	0.57939
f1-score	0.71923	0.16198	0.57939	0.44061	0.68358
support	18099.0	1237.0	0.57939	19336.0	19336.0

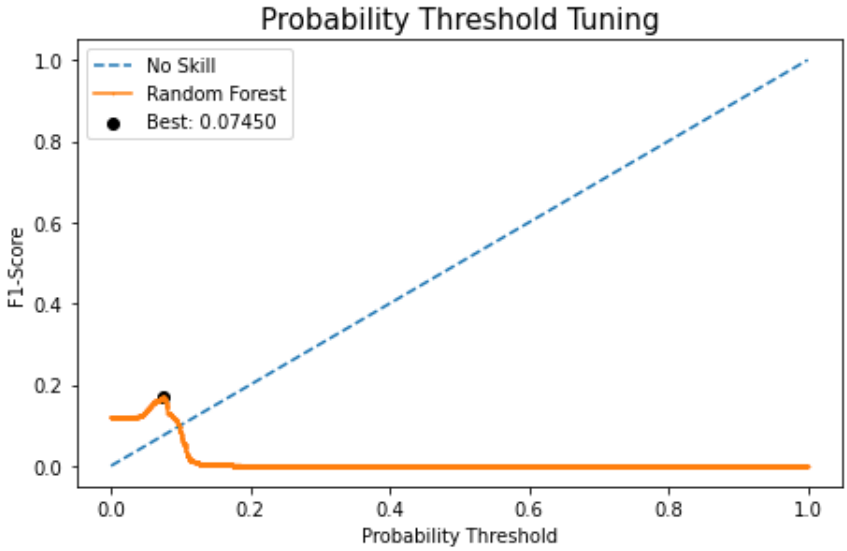
- Precision-Recall Curve - Best Threshold 0.18906, F-Score nan



Confusion Matrix

	0	1	accuracy	macro avg	weighted avg
precision	0.93602	0.0	0.93597	0.46801	0.87614
recall	0.99994	0.0	0.93597	0.49997	0.93597
f1-score	0.96693	0.0	0.93597	0.48346	0.90507
support	18099.0	1237.0	0.93597	19336.0	19336.0

- Threshold Tuning - Optimal threshold 0.0745, F1-score 0.16929



Confusion Matrix

	0	1	accuracy	macro avg	weighted avg
precision	0.9502	0.10619	0.73811	0.5282	0.8962
recall	0.76004	0.41714	0.73811	0.58859	0.73811
f1-score	0.84455	0.16929	0.73811	0.50692	0.80135
support	18099.0	1237.0	0.73811	19336.0	19336.0

- The threshold tuning method performs better and gives good f1 score for class 1.
- Final Competiton Rank is 38
  - Public LB -- Rank: 61 Score: 0.1671883631
  - Private LB -- Rank: 38 Score: 0.1749408983