SUMMARY:

We created the prototype of a framework for transparent model evaluation and explanation. Our project aggregates current open source tools such as SHAP, ELI5, Lime. These libraries provide a various set of visual explanations for trained models. For demonstration purposes, our team created different classification models for “Black Friday” and “Bank Loan Status” datasets. We defined a standard pipeline which consists of four steps and can be easily reproduced for other classification tasks.

Український Католицький Університет | Applied Sciences Faculty

DATA SCIENCE TRANSPARENCY FRAMEWORK:

LIGHT IN THE DARKNESS OF COMPLEXITY

by

Perepichka, Lut, Sirskiy & Riazantsev

1 DATA

The developed framework was validated on “Black Friday” and “Bank Loan Status” datasets. However, the implemented tool was designed to work with any preprocessed data.

2 PREPROCESSING

Data specific cleaning and feature engineering required to use our tool in most cases require classical data preprocessing steps:
1. Removal of duplicated rows
2. Adding dummy variables instead of categorical variables

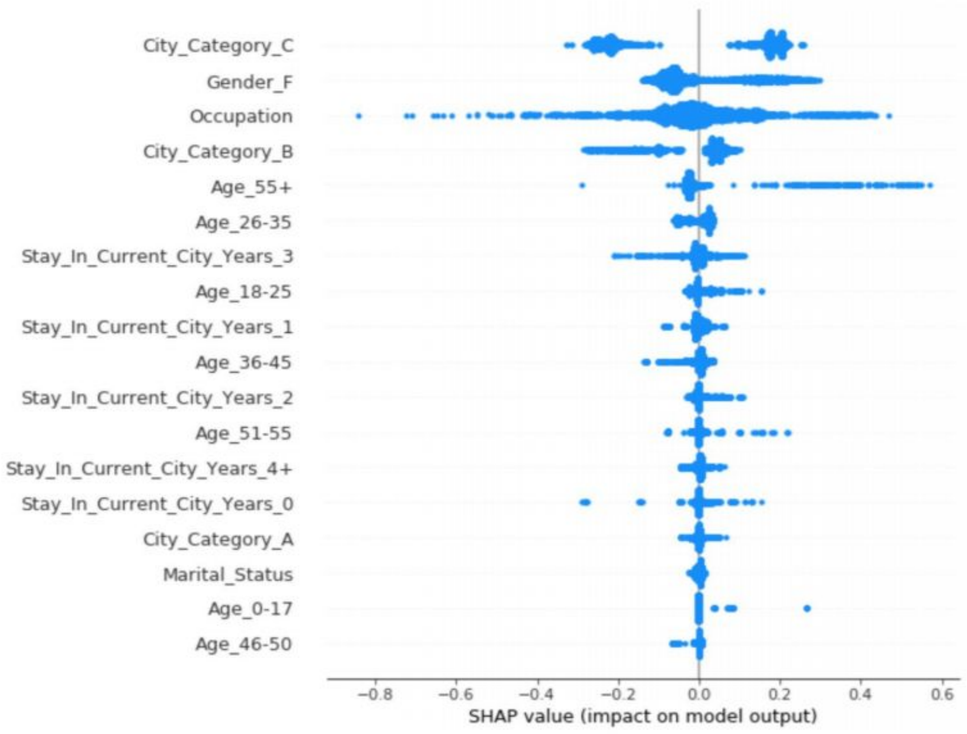
3 MODELLING

The tool requires users to build and train models. Our API will chain them into the common pipeline to use their output in the next step.

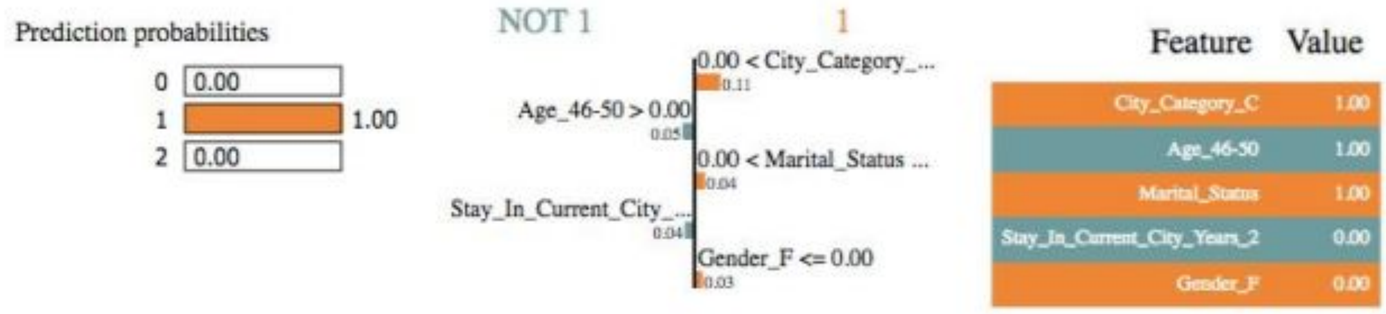
4 EXPLANATION

Our tool aggregates explanations of SHAP, LIME, and ELI5 tools over the data and provided models.
This results can be received as HTML5 outputs.

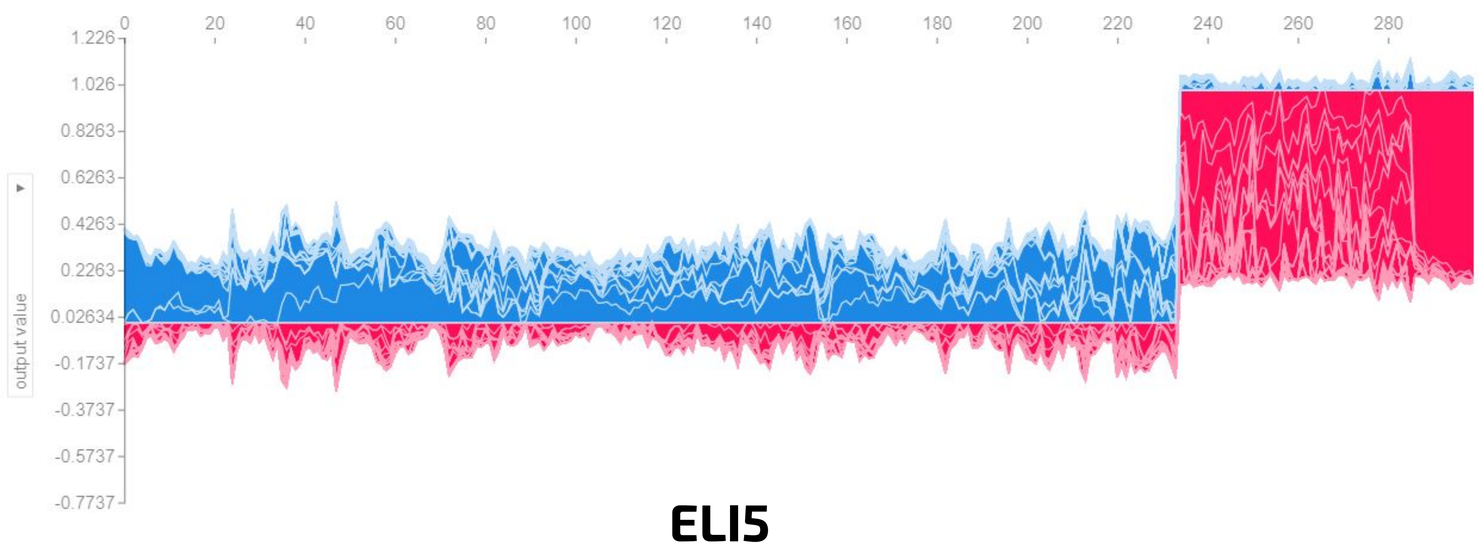
SHAP



LIME



SHAP



ELI5

y=0 (probability **0.343**, score **-0.057**) top features		y=1 (probability **0.415**, score **0.133**) top features		y=2 (probability **0.243**, score **-0.402**) top features	
Contribution[?]	**Feature**	**Contribution**[?]	**Feature**	**Contribution**[?]	**Feature**
+0.180	City_Category_C	+0.120	City_Category_C	+0.137	Gender_F
+0.044	City_Category_B	+0.044	Occupation	+0.107	Age_26-35
+0.020	Stay_In_Current_City_Years_3	+0.033	<BIAS>	+0.026	Marital_Status
+0.011	Marital_Status	+0.023	Age_26-35	+0.012	Age_36-45
+0.009	City_Category_A	+0.015	Stay_In_Current_City_Years_0	+0.003	Age_55+
+0.006	Stay_In_Current_City_Years_2	+0.007	Age_36-45	+0.001	Stay_In_Current_City_Years_2
+0.004	Age_36-45	+0.004	Stay_In_Current_City_Years_1	+0.001	Stay_In_Current_City_Years_0
+0.002	<BIAS>	+0.001	City_Category_B	-0.002	Occupation
+0.001	Age_46-50	+0.001	Age_0-17	-0.003	Age_46-50
-0.002	Age_18-25	+0.001	Stay_In_Current_City_Years_4+	-0.009	Stay_In_Current_City_Years_3
-0.006	Stay_In_Current_City_Years_4+	-0.000	Age_51-55	-0.028	City_Category_B
-0.006	Stay_In_Current_City_Years_0	-0.000	Age_55+	-0.028	City_Category_A
-0.008	Age_0-17	-0.002	City_Category_A	-0.036	<BIAS>
-0.014	Age_55+	-0.004	Gender_F	-0.102	Stay_In_Current_City_Years_1
-0.026	Gender_F	-0.006	Age_18-25	-0.482	City_Category_C
-0.031	Age_51-55	-0.011	Stay_In_Current_City_Years_3		
-0.039	Age_26-35	-0.018	Stay_In_Current_City_Years_2		
-0.046	Stay_In_Current_City_Years_1	-0.024	Age_46-50		
-0.156	Occupation	-0.052	Marital_Status		

SUMMARY

The provided framework gives an ability to see the results produced by ELI5, Lime, and SHAP tools, cross-validate calculated significances and understand whether the model has any serious issues. The code is implementing an API which allows extending the current solution quickly. Adding new datasets, models, or other explanation frameworks does not require architectural changes to the project and can be performed without any struggle.