# ML Assessment 5

Rodrigo Velázquez E.

Leandro Sartini de Campos

Mohammed Maaz

Aman Nain

Prakash Pun

Saraka Sahaswin

Bibek Ranjit

AI & DS, Loyalist College of Toronto

M07 Group 1: AISC1003 - Machine Learning I

Prof. Bhavik Gandhi

Churn Prediction

Contents

[**ML Assessment 5 1**](#_gjdgxs)

[Recap from Week 11 Assignment 3](#_30j0zll)

[Sensitivity Analysis on Linear Regression 3](#_546tsnw6zok2)

[Interpretation of SVM (LIME and SHAP) 8](#_1fob9te)

[LIME 8](#_vrxynp3lzvks)

[Local Feature Importance 8](#_pj3t6r26kj1b)

[Local Instance 1 9](#_wq2xd5auct94)

[Local Instance 2 9](#_qyh9chnbnbs8)

[Local Instance 3 10](#_kah5vbdfzd24)

[SHAP 10](#_3wsmfkbf95is)

[Global Feature Importance 11](#_3znysh7)

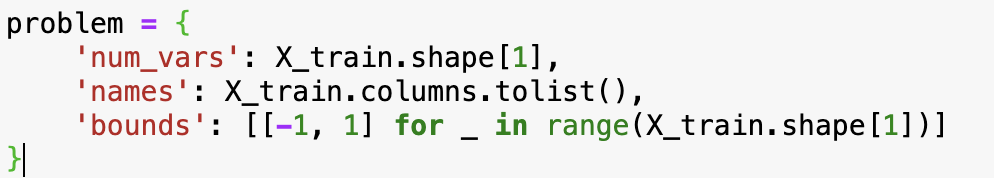
[Findings/ Conclusion 12](#_2et92p0)

[MECE Table 12](#_tyjcwt)

## Recap from Week 11 Assignment

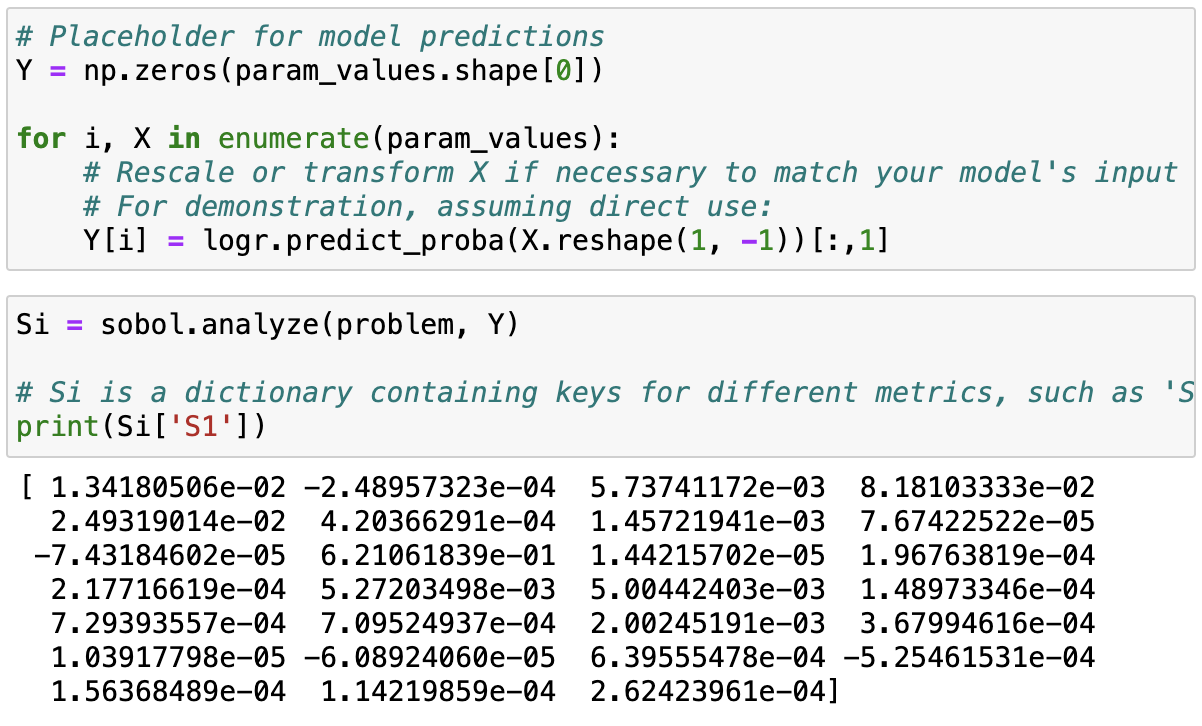
1. EDA
2. Correlation
3. One-Hot Encoding
4. Scaling
5. Balance Data
6. logistic regression and SVM model
7. Built a confusion matrix and AUC plot for the model

## Sensitivity Analysis on Linear Regression

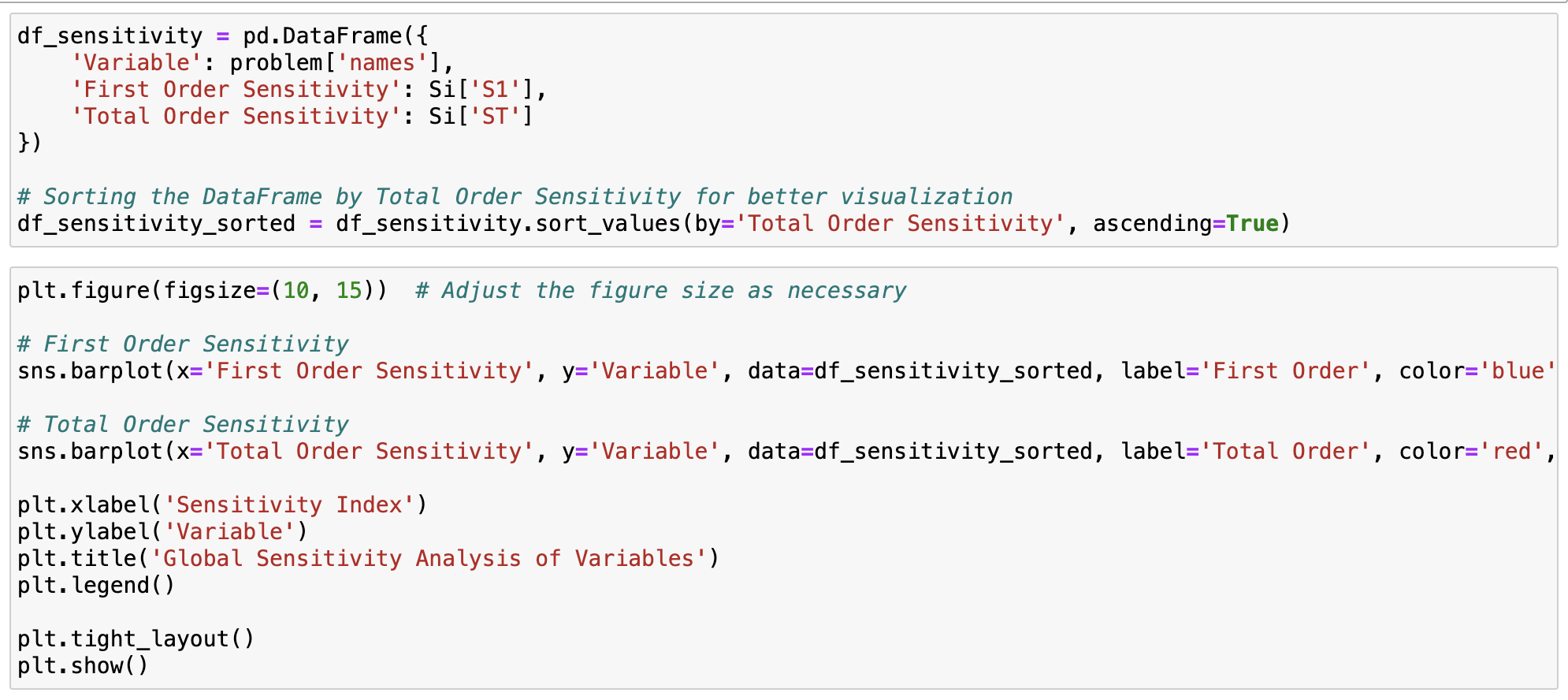
1. We define a problem specification for sensitivity analysis, specifying the number of variables, their names, and their bounds.

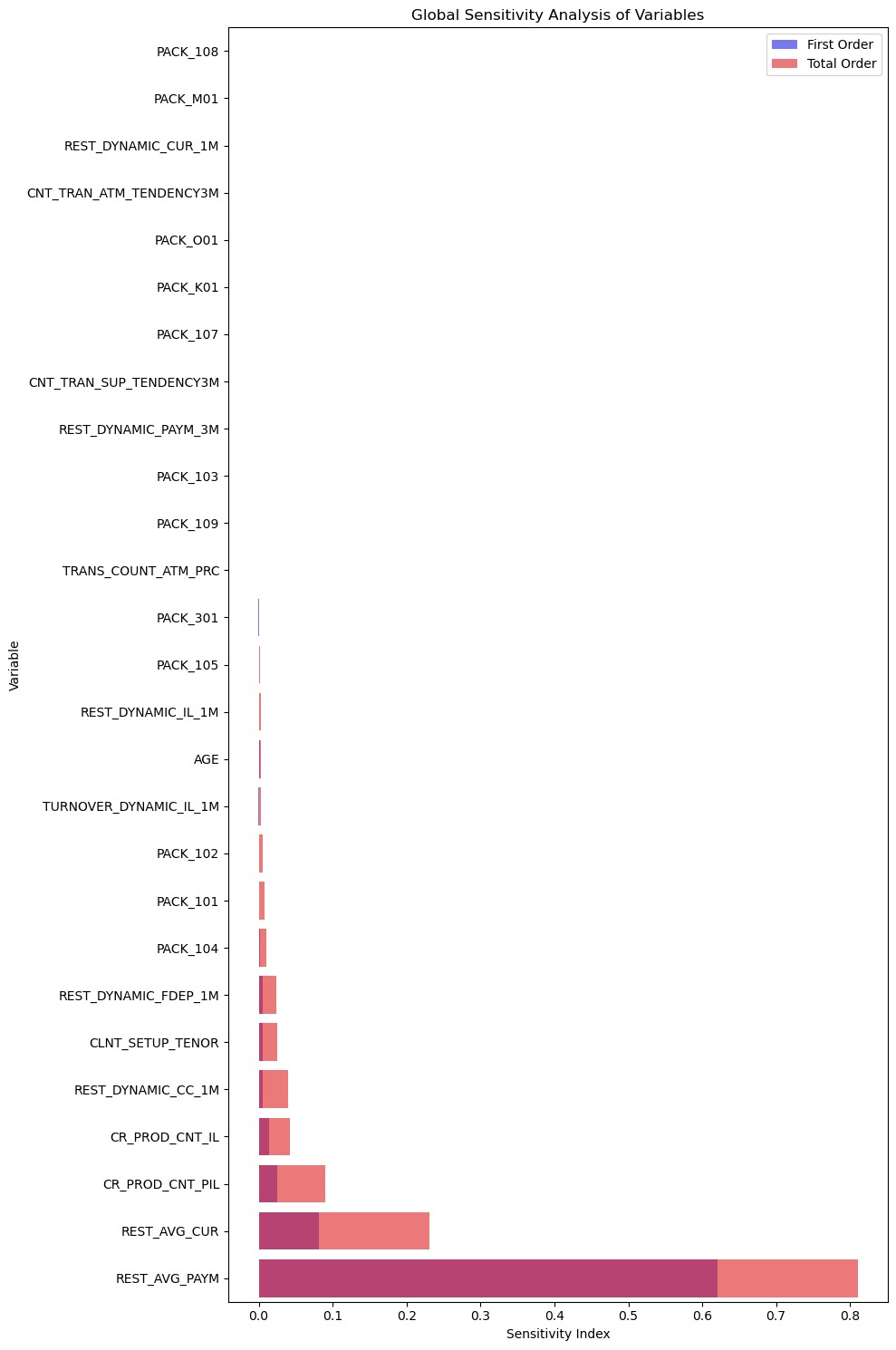
* ‘num\_vars’: specifies the number of variables in the sensitivity analysis problem.
* ‘names’: List of variable or column names
* ‘bonds’: Specifies the bounds (ranges) for each variable in the sensitivity analysis.

1. Generate input samples using the Sobol sampling where each sample represents a combination of input parameters (values for each variable) that will be used to evaluate the model for sensitivity analysis.
2. Evaluate our logistic regression model for each set of input parameters.
3. Performed Sensitivity analysis using SALib’s ‘sobol.analyze()’ function.

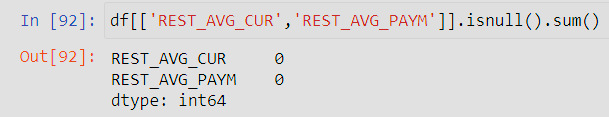


* ‘S1’: represents the first-order sensitivity indices, which tells the individual effects of input variables on the model output.
* ‘ST’: represents the total-order sensitivity indices, which measure the total effect of the input variable on the model output.
* ‘S2’: represents the second-order sensitivity indices, which tells the pairwise interaction between two input variables.
* ‘sensitivity\_conf’: contains the confidence intervals for the sensitivity indices computed.

1. We then create a ‘df\_sensitivity\_sorted’ variable with input variables ordered based on the total-order sensitivity, from least to most influential. And then plot the visual representation of the sensitivity analysis result, allowing for easy comparison of the first-order and total-order sensitivity indices for each input variable.



As we can see most variables have low sensitivity indices, however, “REST\_AVG\_CUR” and especially “REST\_AVG\_PSR” show significantly higher sensitivity which means these variables have a greater impact on the model output.



And when we check this on the original Data frame, it’s not a variable that we treated, but it may be a variable that creates other variables.

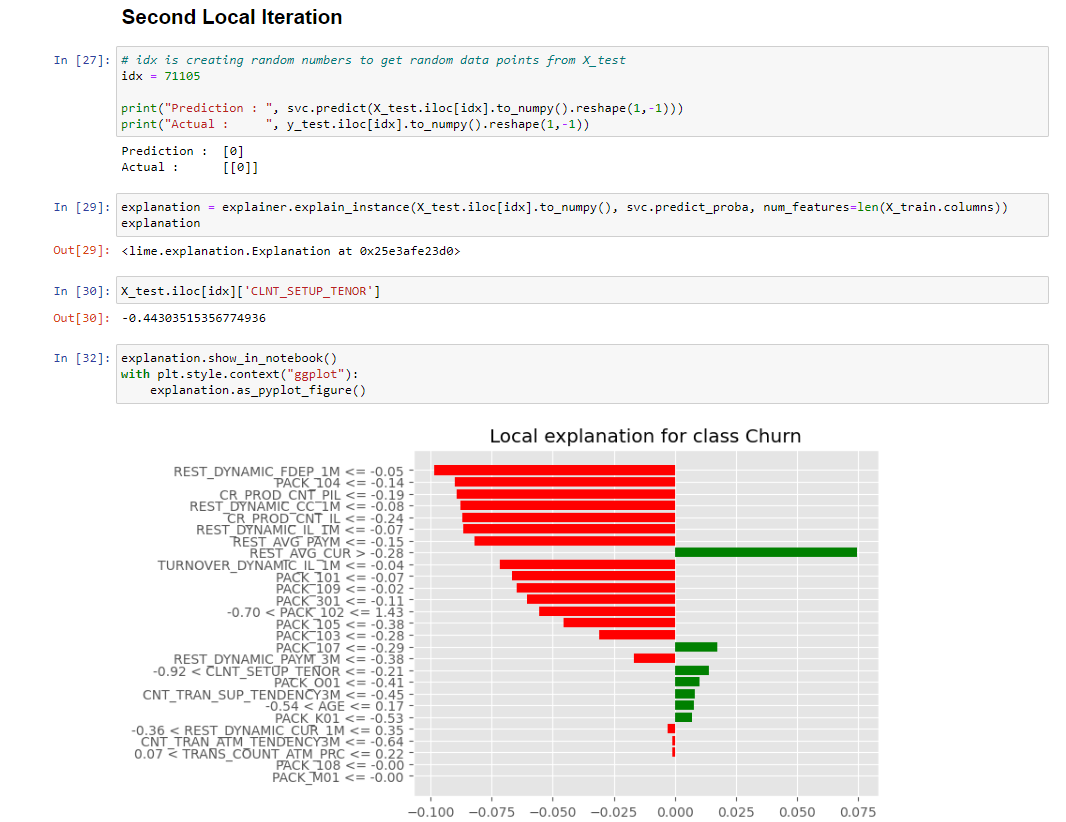
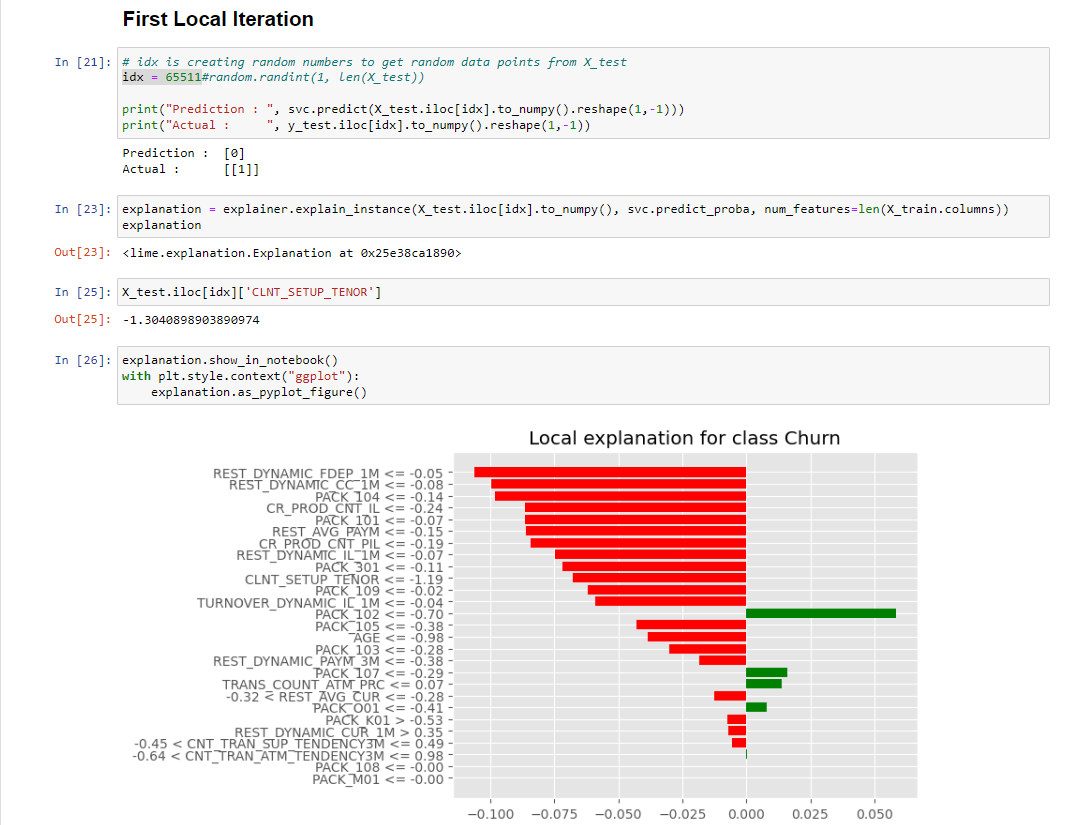
## 

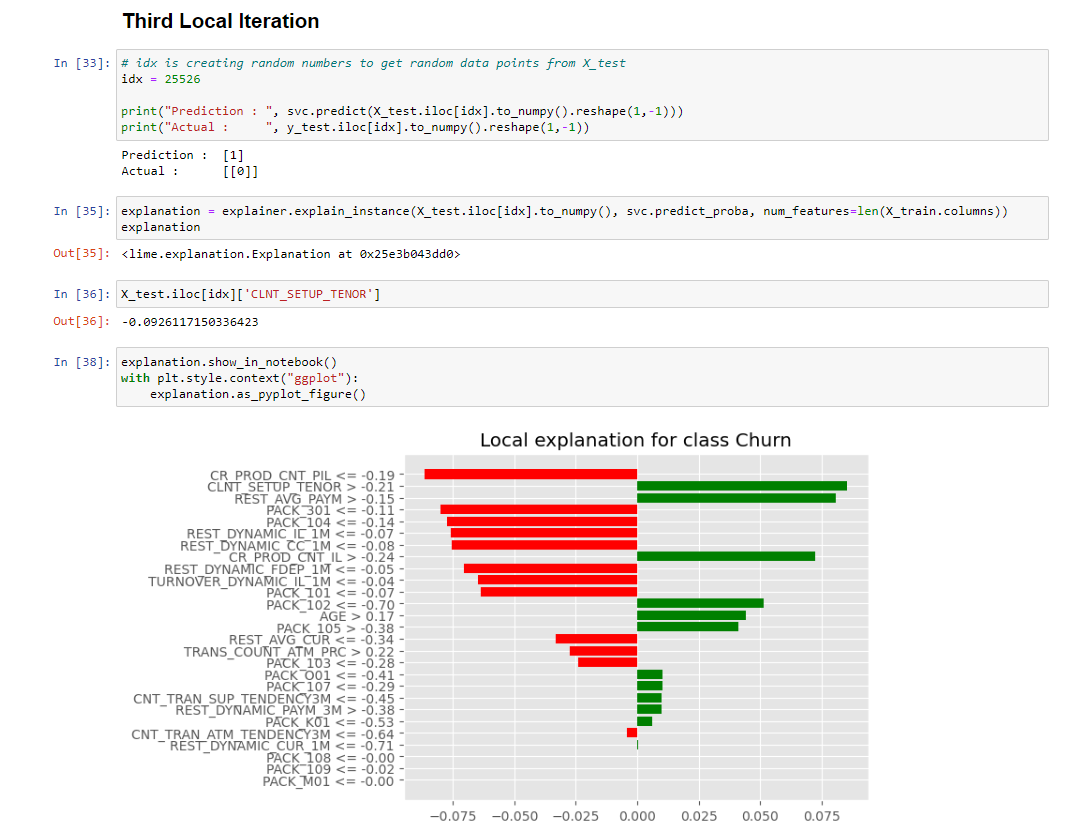
## 

## Interpretation of SVM (LIME and SHAP)

### LIME

We are generating values for idx with random.randint to run LIME on those specific instances as LIME can only be used for individual predictions.

Following are 3 instances generated through LIME.



### 

### 

### Local Feature Importance

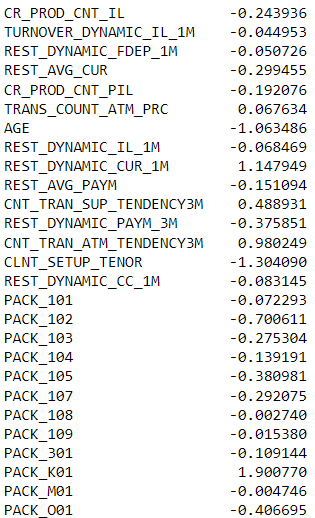
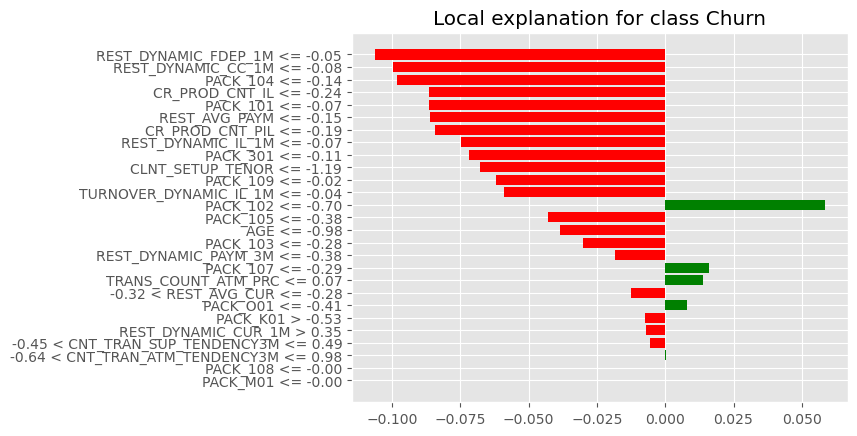
Out of the above three instances, we can see that in ⅔ scenarios, feature PACK\_104(One Hot encoded from a given feature) has caused the prediction to move towards 0. While features like PACK\_107 contribute to the first two instances positively, the contributions of other features have moved it to the negative side.

In the 3rd instance, as the prediction is positive we can clearly observe features which have contributed positively for the model to predict 1.

#### Local Instance 1

*idx = 65511*

*Prediction : [0] Actual : [1]*

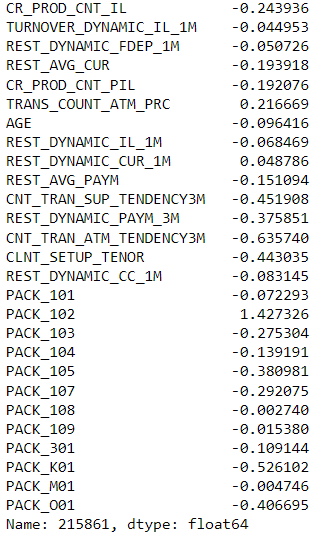
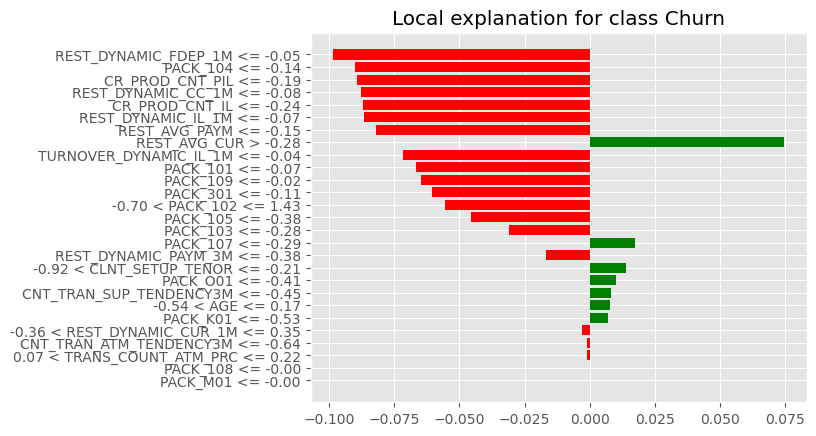


In this datapoint we can see that, while values like REST\_DYNAMIC\_CUR\_1M,CNT\_TRAN\_ATM\_TENDENCY3M have higher values, they have lesser significance in the model affecting the prediction. While features like PACK,REST\_DYNAMIC\_PAYM\_3M have affected prediction positively, due to the high weightage towards the negative side, the model has predicted negative.

#### Local Instance 2

*idx = 71105*

*Prediction : [0] Actual : [0]*

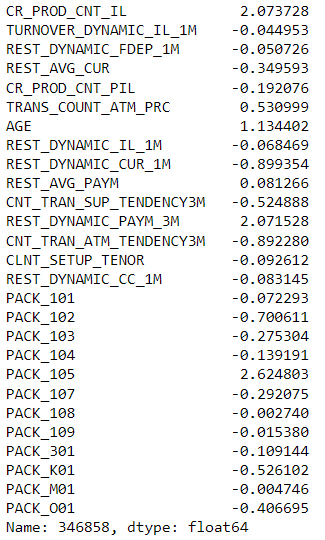
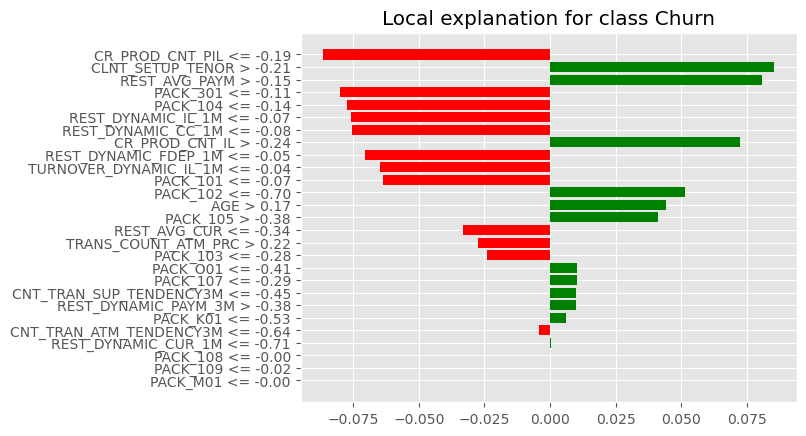


In this instance, the model has predicted correctly as 0. While the value for REST\_AVG\_CUR is -0.193918, looking at this graph we can infer that, for the model this feature is heavily weighted and positively affected to the target. However as the previous instance, as the negative affected features are higher and have more weight, the model has predicted negative.

#### Local Instance 3

*idx = 25526*

*Prediction : [1] Actual : [0]*



In this instance, while CR\_PROD\_CNT\_PIL has the highest weight and is negative, there are multiple features which are affecting the model positively. As the prediction is false we can infer that the weights for the positive might’ve been calculated at a glance, however further investigation is required for an informed decision.

### SHAP

SHAP (SHapley Additive exPlanations) is used to explain machine learning model output. Each feature's contribution to the model's prediction is indicated, providing both local and global interpretability. SHAP values, which are based on cooperative game theory, offer consistent and unambiguous explanations for why a model generates particular predictions, which helps with debugging and understanding of models in a variety of contexts.

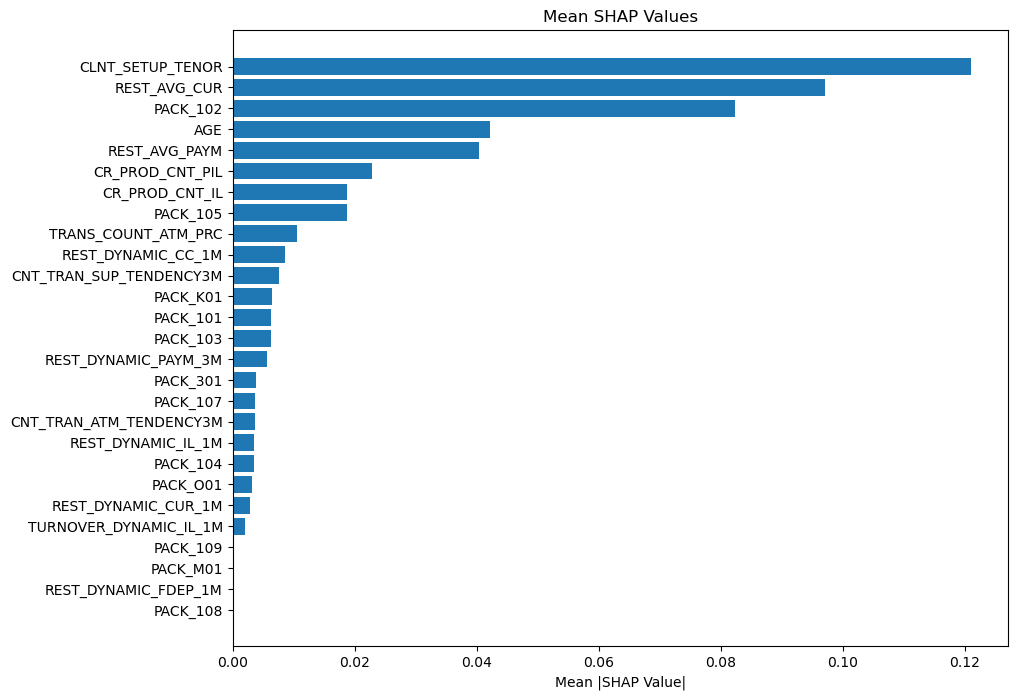


### Global Feature Importance

Global feature importance with SHAP refers to the overall contribution of each feature across the entire dataset in influencing the model's predictions. When analyzing feature importance, SHAP (SHapley Additive exPlanations) values offer a thorough insight by taking feature interactions and their effects on model outcomes into account. These values can be analyzed to see which features have the biggest impact on the model's predictions. This information can then be used to modify the model, choose features, and gain a better understanding of the underlying patterns in the data.

In our dataset, We have used a subsample of 100 data points to calculate the mean SHAP values. We are taking the absolute of the said SHAP values to calculate the mean SHAP, so that negative and positive values don't cancel each other.

By the following diagram we can infer that features like CLNT\_SETUP\_TENOR(months being a customer),REST\_AVG\_CUR (Average current account balances),PACK\_102(Clients with 102 as severance package) and AGE have the most weightage in the model and more likely to decide the outcome/ prediction of the model.



## 

## Findings/ Conclusion

Previously, results were not good as the AUC curve was below the middle line. Even after converting age from months to year, there was no change in error. We checked the PACK column as it was impacting positive/negative classes for CHURN using Lime. The PACK column was removed completely to check for the values that could be biasing the model. Error was the same.

After seeing the data, the PACK class had very imbalanced data. So, we stratified our train test validation creation. We did get better results in comparison to previous ROC AUC. We saw that there are few outliers in different columns. Treating the outliers did not seem to improve the model. Though we got a better ROC Curve, but yet precision is not good. We are slowly progressing

## MECE Table

|  |  |
| --- | --- |
| **Code Quality** | Mohamed Maaz: Did the code quality and comments Aman Nain: Did the code quality |
|  |
| **Sensitivity Analysis** | Rodrigo Velazquez: Did the Sensitivity Analysis Aman Nain: Helped with the Analysis |  |
|  |
| **Local feature importance** | Leandro Sartini: Performed The local feature importance using Lime Mohamed Maaz: Helped with the analysis |  |
|  |
| **Global feature importance** | Saraka Sahaswin: Did the global Feature Importance Prakash Pun: Did the global feature importance Bibek Ranjit: Did the feature importance |  |
|  |
| **Model Tuning** | Leandro Sartini: Did the Model Tuning Rodrigo Velazquez Did the Model Tuning |  |
|  |
|  |
| **Report** | Bibek Ranjit: Did the Report Prakash Pun: Did the Report Saraka Sahaswin: Did the report |  |
|  |
|  |