Lapse Drivers

Report guide

Janco Van Niekerk

2024

Contents

[1. Introduction 2](#_Toc177490263)

[2. Correlations 2](#_Toc177490264)

[2.1. Shapley Values 3](#_Toc177490265)

[2.2. Feature Importance 3](#_Toc177490266)

[2.3. Feature Effects 4](#_Toc177490267)

[3. Causal Analysis 6](#_Toc177490268)

[3.1. Net Risk Reduction 6](#_Toc177490269)

[3.1. Variable Alteration Methods 7](#_Toc177490270)

[3.2. Non changeable variables 9](#_Toc177490271)

[4. Conclusion 10](#_Toc177490272)

List of Figures:

[Figure 1 - Shap Bar Plot 3](#_Toc177490234)

[Figure 2 - Shap Beeswarm Plot 4](#_Toc177490235)

[Figure 3 - Shap Scatter plot for Claim Count 4](#_Toc177490236)

[Figure 4 - Shap Scatter plot for Economy health 5](#_Toc177490237)

[Figure 5 - Shap Scatter plot for Home Language Venda 5](#_Toc177490238)

[Figure 6 - Shap Scatter Plot for Newsletter Email Count 5](#_Toc177490239)

[Figure 7 - Causal Graph Example 6](#_Toc177490240)

[Figure 8 - Net Risk Reduction of uniform Document Pages 8](#_Toc177490241)

[Figure 9 - Net Risk Reduction for change in Current Document Pages 8](#_Toc177490242)

[Figure 10 - Net Risk Reduction of Increasing Current Newsletter Emails 9](#_Toc177490243)

# Introduction

This project aims to develop a predictive model and data pipeline to identify the drivers of customer lapses, which will inform retention campaigns. The model and pipeline are designed to handle unspecified data fields and sparse survey data.

As mentioned, the primary objective is to uncover and analyse the drivers of customer lapses. To achieve this, we will use interpretability tools, such as SHAP, to reveal key relationships between input features and predicted outcomes. However, it's important to note that these tools highlight correlations rather than causations. As discussed in the Microsoft notebook [here](https://github.com/shap/shap/blob/master/notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.ipynb), correlations identified by these techniques do not imply causation.

To accurately assess the impact of business changes on specific variables, we must develop causal models. This requires making certain assumptions and applying tools from causal analysis.

The project generates several outputs that are essential for reporting purposes. This document explains how to correctly interpret these outputs from a business perspective and outlines how to implement them within a software context.

During the development phase, a list of dictionaries was created and saved as a .pkl file. This file contains various key details that are required for reporting.

The .pkl file can be imported into a Python script and used in several ways:

* Integrate the data into a processing pipeline, allowing further transformations before saving it to a database. Once stored, the data can be accessed and utilized by multiple applications.
* Import the data into a Python application, such as a Django or Streamlit app, to support various functions like displaying graphs on a dashboard.

The script ‘Report\_Processing\_Example.py’ demonstrates how to process the .pkl file.

Some of the processed feature names in the .pkl file may not be intuitive for business stakeholders, as they include technical preprocessing terms. In a report, application or database, these feature names can be adjusted to more user-friendly terms that are familiar to the business.

# Correlations

The calculated Shapley values can be used to analyse the relationships between different features and the probability of customer lapses. This provides valuable insights for the business and helps in forming hypotheses about why these relationships exist. These insights also lay the foundation for applying more advanced causal analysis techniques, which are discussed later in this document. The Shapley values can be accessed using the ‘shap\_values’ key in a dictionary item of the .pkl file.

If available, recommended feature names are listed under the ‘shap\_feature\_name\_changes’ key. Additional feature name adjustments can also be made in the ETL pipeline, report, or application as needed.

## Shapley Values

Shapley values can be used to provide both global explanations (overall model behaviour) and local explanations (why the model made specific predictions).

Higher Shapley values indicate an increase in the probability of customer lapse, while lower values suggest a decrease in lapse probability for a given feature.

## Feature Importance

The mean absolute Shapley values help to understand the overall (aggregated) contribution of each feature across all customers in the sample. This makes it a useful metric for assessing the importance of each feature.

A Shap bar plot can be used to visualize these mean absolute Shap values. An example of this is given in Figure 1.

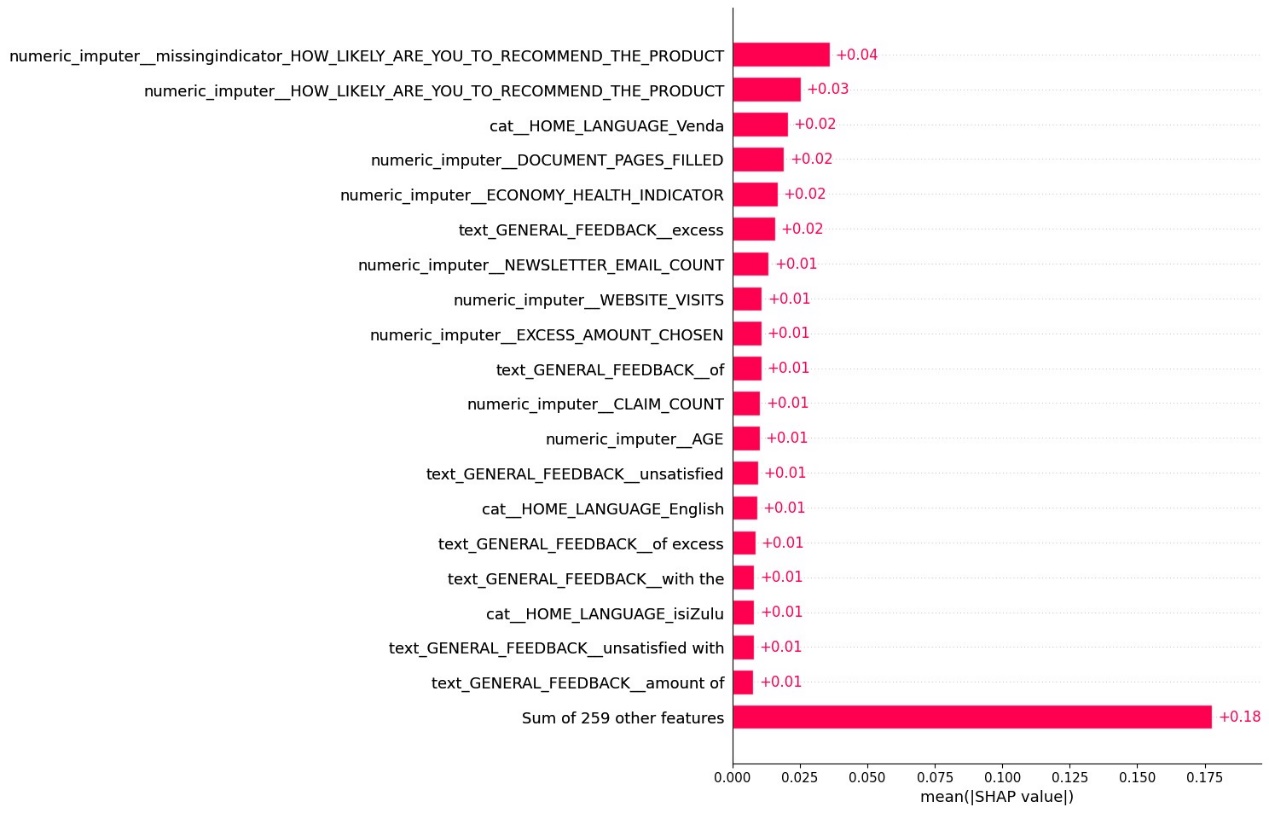


Figure - Shap Bar Plot

## Feature Effects

Beeswarm plots were used to visualize the effect (or correlation) of each feature on lapse probabilities for both high and low feature values. This offers a clear overview of how different feature values influence the probability of customer lapse across various features.

An example of such a Beeswarm plot is given in Figure 2.

To gain a more detailed understanding of how specific feature values correlate with lapse probability, scatter plots are used. These plots display the relationship between feature values and lapse probability, helping to identify patterns and trends.

Examples of these scatter plots are included in Figure 3, Figure 4, Figure 5 and Figure 6.

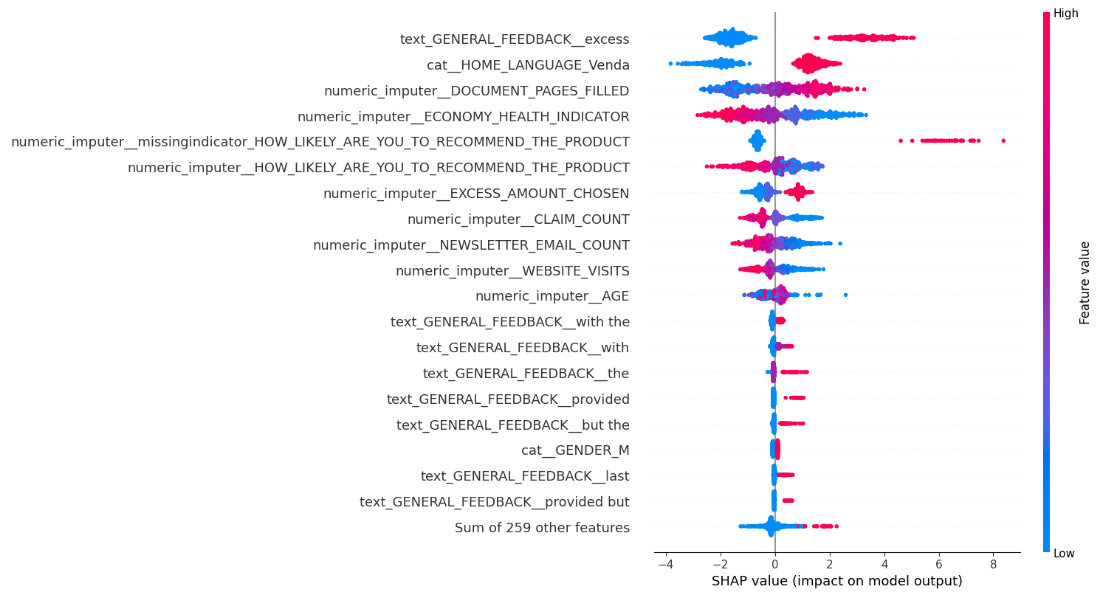


Figure - Shap Beeswarm Plot

A graph showing the number of numbers

Description automatically generated

Figure - Shap Scatter plot for Claim Count

A graph showing the number of numbers

Description automatically generated with medium confidence

Figure - Shap Scatter plot for Economy health

A graph with blue dots

Description automatically generated

Figure - Shap Scatter plot for Home Language Venda

A graph showing the number of numbers

Description automatically generated with medium confidence

Figure - Shap Scatter Plot for Newsletter Email Count

# Causal Analysis

While SHAP explanations are very useful, they do not establish causality. As mentioned earlier, you cannot directly infer from SHAP values alone that "changing X will decrease lapse probability."

To estimate the causal effects of changing variables, it’s crucial to work with business domain experts and develop a hypothesized causal graph (see Figure 7 for an example). Shapley value plots can be used as a starting point to guide discussions with these experts and help in creating the causal graph.

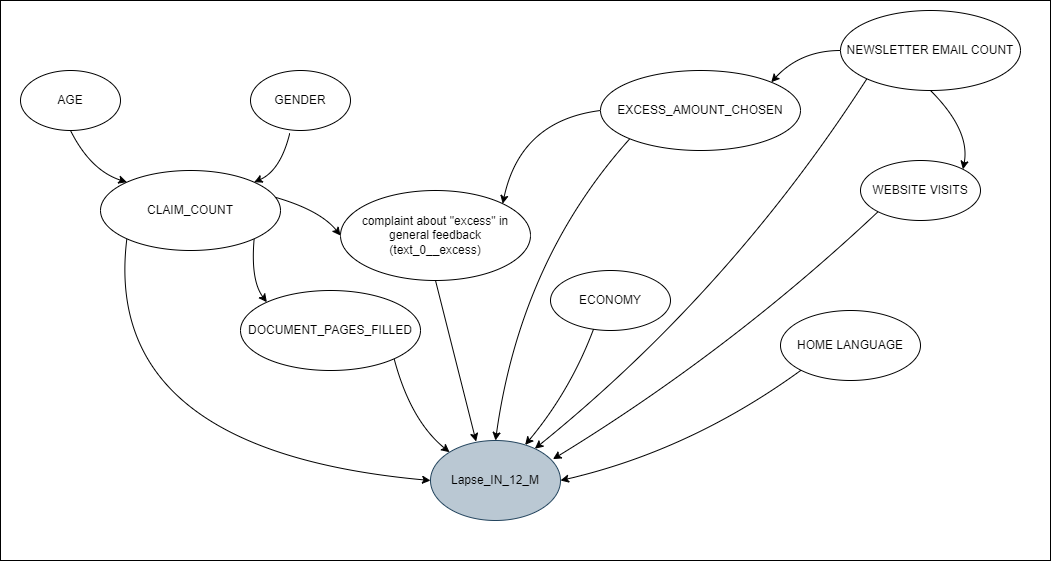


Figure - Causal Graph Example

With the hypothesized causal graph, we can apply appropriate causal analysis techniques to estimate how changing specific variable values will affect lapse probability.

Often, modifying a variable will involve a business action, which may incur costs. Therefore, we calculate the 'net risk reduction' relative to the cost of implementing these actions to retain customers.

## 3.1. Net Risk Reduction

When using causal inference techniques (such as the techniques discussed in the next section), simply stating the change in lapse probability may not give stakeholders full understanding of the business outcomes associated with changing the treatment value. It is for this reason that we need to investigate a more informative approach.

Changing the value of the treatment may have a cost attached to it. Therefore, it will be useful to know what the gain (or ‘net risk reduction’ in the case op lapse reduction) of changing the treatment value is so that we can compare it with the cost and make the business decision on whether it is worth a change (i.e. changing the treatment value) being implemented.

The change in risk for a single customer subjected to a different treatment can be defined as follows:

Risk\_Change = Lapse\_impact\*△Lapse\_probability

Where:

* Lapse\_impact: the cost associated with losing the customer i.e. the revenue which you would not receive from them for the next 12 months.
* △Lapse\_probability: Change in the probality of lapse if subjected to the new treatment

Therefore, using the △Lapse\_probability of each customer, the total risk change for all customers in the sample can be calculated as:

Total\_risk\_change = (Lapse\_impact \* △Lapse\_probability\_1) + (Lapse\_impact \* △Lapse\_probability\_2) + ..... + (Lapse\_impact\*△Lapse\_probability\_n)

If we assume that the lapse impact is the same for all customers, we get:

Total\_risk\_change = Lapse\_impact\*(Sum\_of\_all\_△Lapse\_probabilities)

To get the total reduction in risk, we do the following:

Total\_risk\_reduction = (-1\*Total\_risk\_change)

As mentioned, it should be determined if the cost associated with implementing a change is worth it. This can be determined by calculating the 'Net of risk reduction':

Net\_risk\_reduction = Total\_risk\_reduction - Implementation\_Cost

Because we export the .pkl file, stakeholders will have the option to change the ‘Implementation cost’ and ‘Lapse Impact’ if required.

## 3.1. Variable Alteration Methods

We will calculate the causal effect of changing certain values on net risk reduction by using the techniques discussed in the following sections.

Depending on business needs, we will evaluate the effect of changing variables in two main ways:

1. Change all values of the treatment to the **same uniform** value: As an example, lets propose that certain stakeholders mentioned that they think it is possible to change the paperwork process so that each of their customers only ever have to fill out documentation once and that they believe these documents can be the same amount of pages for everyone. As an example, Figure 8 shows that there is a net risk reduction of about half a 300 000 Rand if each customer only fills out 8 pages of documentation.
2. Increase/decrease the **current** treatment values by a specific amount: Other stakeholders believe that it will not be possible to create a single and similar paperwork process for all their customers. They argue that certain customers are unique and therefore require additional document pages to be filled. They are convinced however, that they can reduce everyone’s document pages by 3 pages ex. if a customer usually fills 10 pages, they will now only fill 7. To illustrate this point, we can see from Figure 9, that there is a net risk reduction of about R150,000 if the amount of paperwork for each customer is decreased by 2 pages from their current amount of paperwork. As another example Figure 10 shows that the net risk reduction of sending an additional 10 email newsletters will amount to an estimated R400,000.

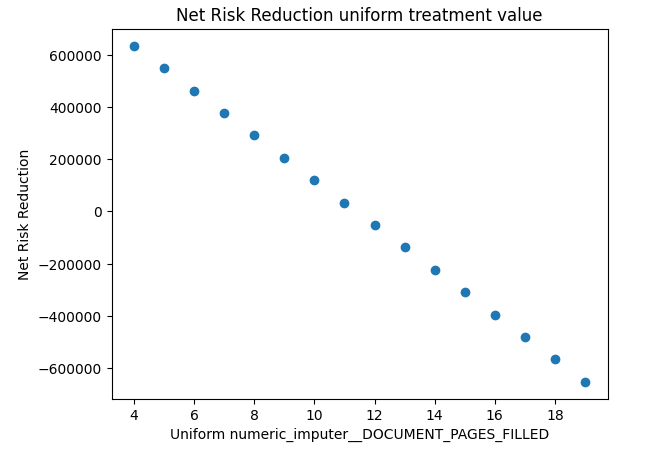


Figure - Net Risk Reduction of uniform Document Pages

A white sheet with blue dots

Description automatically generated

Figure - Net Risk Reduction for change in Current Document Pages

A graph with blue dots

Description automatically generated

Figure - Net Risk Reduction of Increasing Current Newsletter Emails

## Non changeable variables

Certain variables cannot be changed. However, if we do see a correlation between a certain feature of interest and lapse probability, we can hypothesize why this might be occurring.

For example, when looking at the shap values of "HOME\_LANGUAGE\_Venda" (Figure 5) we can see that there is an increase in lapse probability if the customer's home language is Venda (this is where the variable equals 1).

If we study the causal graph and see that this variable is not confounded by others, this will suggest that the SHAP value plot is a good representation of the actual relationship.

Together with business stakeholders, we can hypothesize why this might be, for example:

* Venda is not offered as a customer service language option.

Based on this hypothesis, a suitable business action could be:

* Offer Venda as a language option for customer service.

# Conclusion

This document has outlined the process of using Shapley values and causal inference techniques to understand and address customer lapses. While Shapley values provide valuable insights into feature impacts on lapse probability, they do not establish causality. Therefore, it is crucial to develop a causal graph with input from business domain experts to estimate the true effects of changing variables.

We discussed the importance of assessing the cost-effectiveness of proposed changes by calculating the net risk reduction, which considers both the impact on lapse probability and the cost of implementation. By comparing the total risk reduction with the implementation costs, stakeholders can make informed decisions about whether the proposed changes are worthwhile.

Additionally, we explored methods to address non-changeable variables by hypothesizing potential business actions. Through this approach, we can guide strategic decisions aimed at reducing customer lapses and improving business outcomes.