Table of Contents

[1. Introduction 2](#_Toc69687713)

[2. Business Understanding: Defining the Problem 2](#_Toc69687714)

[Defining the Target Variable 3](#_Toc69687715)

[3. Data Understanding 3](#_Toc69687716)

[4. Data Preparation 3](#_Toc69687717)

[Missing Data Imputation 4](#_Toc69687718)

[Exploratory Data Analysis (EDA) 5](#_Toc69687719)

[5. Modeling 6](#_Toc69687720)

[6. Deployment 7](#_Toc69687721)

[7. References 8](#_Toc69687722)

## Introduction

In the Insurance Industry case study, a combination of predictive modeling and Exploratory Data Analysis approaches are going to be used to improve the efficiency of the Insurance Claims prediction of a large U.S. corporation. Predictive Analytics, on the Insurance Industry, using fictitious company data as a case study. We can analyze the Insurance company’s list of insured buildings to predict which building will make a claim during the insured period of the building and How we can the company leverage this to plan for imminent claims and taking on new buildings.

The basic concept of insurance is that one party, the insurer (the party that takes on the risk), will guarantee payment for an uncertain future event. While another party, the insured, pays a smaller premium (amount) to the insurer in exchange for that uncertain future event. Seeing that taking on risks is the major businesses of insurance companies, having an accurate risk analysis on policies, based on data is crucial to their survival.

Due to earthquake, Flood, Storm, and other issues, we have seen an increase in the number of building collapse in Lagos and major cities in Nigeria. Olusola Insurance Company offers a building insurance policy that protects buildings against damages that could be caused by a fire or vandalism, by a flood or storm.

The problem the Insurance company needed to address is this: Could the company predict the building against damages of problems transcribed from Insurance Claims to predict if a part would be needed to resolve the problem.

Python as a programming language has numerous uses such as Machine Learning, AI, mobile applications, statistical , etc. As we know, AI, machine learning, and data analysis/analytics is where it has amassed most of its popularity. Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc.

## Business Understanding: Defining the Problem

The business objective remained the same: The main goal of the case study is to build a predictive model to predict if a building is going to have an insurance claim during a certain period or not. Predict the buildings against damages related to the building which are insured in Company and we can evaluate how they can predict the building damages actively and prepare for the business model and plans. We will have to predict the probability of having at least one claim over the insured period of the building.

### Defining the Target Variable

The Target variable is going to be Claim. The Claim prediction is going to be 0 or 1. Based on the YearOfObservation, Building Type, we need to predict the Claims.

## Data Understanding

The format of the dataset for this case study is CSV. The data contains below variables.

Pandas library used to load the data from csv to Dataframe.

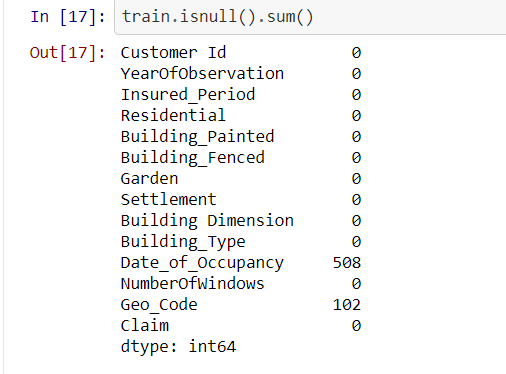
|  |  |
| --- | --- |
| **Variable** | **Description** |
| Customer Id | Identification number for the Policy holder |
| YearOfObservation | year of observation for the insured policy |
| Insured\_Period | duration of insurance policy in Olusola Insurance. (Ex: Full year insurance, Policy Duration = 1; 6 months = 0.5 |
| Residential | is the building a residential building or not |
| Building\_Painted | is the building painted or not (N-Painted, V-Not Painted) |
| Building\_Fenced | is the building fence or not (N-Fenced, V-Not Fenced) |
| Garden | building has garden or not (V-has garden; O-no garden) |
| Settlement | Area where the building is located. (R- rural area; U- urban area) |
| Building Dimension | Size of the insured building in m2 |
| Building\_Type | The type of building (Type 1, 2, 3, 4) |
| Date\_of\_Occupancy | date building was first occupied |
| NumberOfWindows | number of windows in the building |
| Geo Code | Geographical Code of the Insured building |
| Claim | target variable. (0: no claim, 1: at least one claim over insured period). |

## Data Preparation

Based on the reading, “data scientists usually spend 70% of the project time here, preprocessing and exploring the data”. So, the Data Preparation is important for predictive analytics project. In the Case Study project, we need to make sure the data is suitable and well prepared for model. While preparing the Data, we need to verify the metadata of the given dataset and need to understand very clearly about data patterns, relationships etc.

* Need to describe the dataset describe()
* Need to understand the number of columns and rows shape
* Need to understand the summary of the data ( info )

Verify the Null values, these missing values impact the model. Hence, we need to treat these missing data.



### Missing Data Imputation

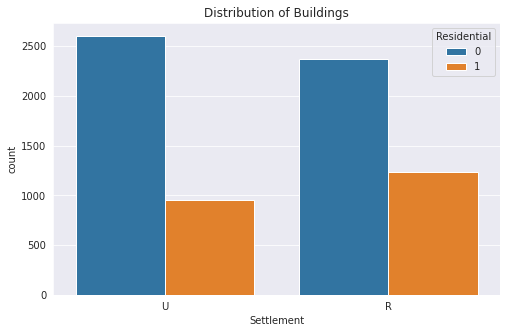
As we know, If the variables are treated as continuous or ordinal, leaving these as 0 communicates to the algorithms that 0 is the best answer and if the data were treated as categorical, these could be left as the value 0.

### Exploratory Data Analysis (EDA)

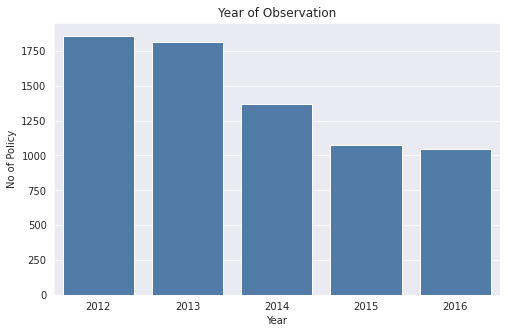
We need to perform EDA and that data needs to be cleaned to gain insights that may be useful in improving the performance of our model. Few of the EDA as follows

* Filling missing information (Null / NA values)
* Encoding of categorical variables
* Dropping of features (or columns)

Visualization used to show behavior of data and explains insight of data. The below BAR chart shows that Residential building is having more settlement.



Date wise policy shown. More Policies were purchased in the year of 2012 and 2015/2016 were less policies sold.



## Modeling

As we know, Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). Hence in our Case Study we can use Logistic Regression methods to predict the outcome claim as 1 or 0 and the data of this case study as categorical data.

* Importing the necessary libraries needed -

from sklearn.linear\_model import LogisticRegression

* Encoding our data –

from sklearn.preprocessing import LabelEncoder

* Splitting our dataset into training and test set –

from sklearn.model\_selection import train\_test\_split

* Perform prediction & Evaluate Model Performance

We need to perform the above steps to perfume and evaluate the model performance. The main four main performance metrics used to evaluate the effectiveness of classification models:

* Accuracy: test’s ability to correctly predict both classes
* Precision: test’s ability to correctly detect positive classes from all predicted positive classes
* Recall (Sensitivity): test’s ability to correctly detect positive classes from all actual positive classes.
* F1 Score: harmonic mean of precision and recall

## Deployment

The individual branches decided to what degree the models were used to influence what changes were made within their branches. The Deployment of Machine Learning or Predictive Analytics Solution, so the deployment will be dependent on how it will be utilized.

## References

1. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die
2. Applied Predictive Analytics by Dean Abbott ( Principles and Techniques for the Professional Data Analyst )
3. Predictive Analytics with Python: Case Study of the Insurance Industry by Emmanuel Ayeni

<https://medium.com/dev-genius/predictive-analytics-with-python-case-study-of-the-insurance-industry-6e2938c2e6fd>

1. https://www.kaggle.com/c/mock-competition-2-ai-funaab/overview