Model Card  
Date: 09/08/2024  
Contents Insurance Risk Analysis   
Author: Victor Acosta

# Model specifications

|  |  |
| --- | --- |
| **Model Type** | Random Forest |
| **Inputs** | EMPLOYMENT  MAR\_STATUS  ALARM  LOCKS  BEDROOMS  NEIGH\_WATCH  OWNERSHIP\_TYPE  PROP\_TYPE  YEARBUILT  LEGALFEES\_COVER  EMERGENCIES\_COVER  KEYREPLACE\_COVER  FLOODING\_RISK  AREA\_RISK |
| **Outputs** | 1 – HIGH RISK  0 – NO RISK |
|  |  |

# Intended Uses

For underwriters generating quotes and policies for property contents insurance.

**Domain and users:** In the UK building and contents insurance is usually divided into the owner of the building (freehold) and the owner of a flat (leasehold), this division provides an excellent market for contents-only insurance. In addition, it can be used for combined building and contents for house owners.

**Out-of-scope**: Any quotes or policies for building-only insurance as the model was not trained for data associated with the building specs.

# Limitations

**Trade-offs.** The model was optimised for detecting high risk policies. Therefore, there may be more cases of high-risk predictions for low-risk scenarios.

**Groups and geographical area**. The main group is UK property market as it has a good division between building and contents products. It may be reused in other countries but perhaps with some accuracy degradation.

**Sensitive data**: Age groups and gender have been removed and does not contribute to the outcome of the predictions

# Datasets

The following inputs have higher rank, and they should be considered mandatory fields to have the best accuracy: **Year built, Property Type, Number of bedrooms, ownership type, emergencies cover and marital status.**

A graph of a bar graph

Description automatically generated

***Figure 1: Feature importance based on a sample of 13104 policies***

A full list and details are provided below:

|  |  |
| --- | --- |
| Feature | Description |
| EMPLOYMENT | R = Retired, E = Employed, O =Other |
| PROPERTY\_TYPE | 1, 2, 9, 10 and 19 |
| OWNERSHIP\_TYPE | 8 or Other |
| MAR\_STATUS | M: Married, P: Partnered, W: Widowed, O: Other |
| AREA\_RISK | 1 – Low Risk, 2 – High Risk |
| ALARM | The property has appropriate alarm |
| LOCKS | The property has appropriate locks |
| NEIGH\_WATCH | Neighbourhood watch |
| YEARBUILT | The year the building was built |
| LEGALFEES\_COVER | Include legal fees |
| EMERGENCIES\_COVER | Emergencies options included |
| KEYREPLACE\_COVER | Replacement of keys included |
| FLOODING\_RISK | Flooding risk of the property based on the surrounding area |

# Evaluation and Results

The original dataset has 147259 insurance policies and can be found [here](https://www.kaggle.com/datasets/xavierdataset/stable-home-insurance-data-driven-growth). This was filtered down to 32761 policies in a ratio of 18019 high and 14742 low risk policies. Finally rebalanced, removed all noise and unnecessary features. The final training and test data was segregated as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Total policies** | **High risk** | **Low risk** |
| Training Data | 13104 | 7215 | 5889 |
| Test Data | 7862 | 4322 | 3540 |
| Validation Data | 6492 | 5313 | 1179 |

Validation data was balanced manually to provide a more realistic scenario where in real life there are only a small number of policies with high risk. However, the other sets were balanced 55/45 as shown above to ensure the model learned how to identify high risk policies.

The accuracy of the model is 76%