Analysis of the occurrence of Events that request an Insurance Claim

Stefanela Stevanović[1]

Abstract

The goal of this project was to analyze the occurrence of events that request an insurance claim, based on the data provided by Zurich Insurance Company. Time analysis has shown that July has more claims than any other month $\sim$ 11.5%, while Monday was the day of the week with the most events $\sim$ 17.8%. Furthermore, the relationship between the number of events for different product groups and country parts was analyzed. Product group 11 had by far the highest frequency of events, especially in the north-west region where 64 events occurred on every 100 underwritten policies.

Keywords
Data analysis, Visualization

[1] *ss51676@student.uni-lj.si, 63220492*

Introduction

The task of project 2 of the Introduction to Data Science course was to provide insight into the data provided by Zurich Insurance Company through text, tables and visualizations. The provided data is divided into three data sets:

- customer data, which contains information about customers, such as age, gender, part of the country, etc.
- claim data that contains the dates of events, claims opening and closing dates, as well as data on the amount of the claim
- policy data, which contains information on types of product, status, policy underwritting and cancellation dates, etc.

These three data sets are connected using foreign keys, such that primary key of policy dataset (policy_id) is a foreign key in claim dataset and primary key of a customer data set (customer_id) is a foreign key in policy and claim data sets.

This project is mainly centered around the analysis of patterns of occurrence of events, in terms of connections between the number of events and different types of products, months of the year and different parts of the country which policy customers come from. Knowledge about these event occurance patterns enables the company to predict costs, but also to spot some irregularities that could potentially stem from the fraud.

Methods

This part of the project describes the way in which the data was prepared and analyzed. This project can be divided into three parts, i.e. three questions that are sought to be answered. Those questions are presented here, as well as the methodology that was used to answer them.

Data Cleaning

Primarily, the rows that do not have customer_id and policy_id information have been removed, as these are key elements for connecting the tables. Since there only 23 claims from the 90s, the rows referring to these years were also removed, and only data from 2000 onwards was used in the analysis. Therefore, the analysis covers a total of 21 years, from 2000 to 2020, including 2020, as the last year for which the data was available in the provided data sets. Columns that will not be part of the analysis have also been removed from the tables, for better visibility.

In which months and days of the week do events happen most often?

In order to answer this question, the number of events that occurred in each month was calculated, as well as the total number of events in the period from 2000 to 2020. The percentage of events that occurred in each month was obtained by dividing the number of events in each month by the total number of events. The same methodology was applied for the days of the week. The number of events that occurred on each day of the week was calculated and divided by the total number of events .

How does month of the year affect the number of events happening for different product groups?

A total of 159 product groups appear in the entire database, numbered from 0 to 158. Since it would be impractical and visually unappealing to create visualizations containing information for all 159 products, the 30 best-selling products were singled out to be used in the analysis. These 30 product groups are shown in Fig.1, as well as the total number of underwritten policies for each of them.

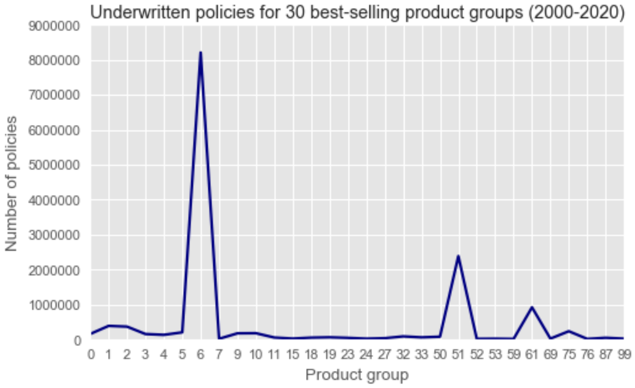

Figure 1. Total number of underwritten policies for 30 best-selling product groups in the period from 2000 to 2020. This figure shows that by far the most policies were sold for the product group 6, followed by products 51 and 61.

For each of the 30 products, the number of events that occurred in each month of the year in the period from 2000 to 2020 was calculated. For the calculation, it was necessary to merge the claim and policy data sets via a policy_id foreign key. However, since the products with the most policies will naturally have the most events, normalization was performed, so that the number of events for each type of product in each month was divided by the total number of underwritten policies for that product group.

Does the number of events per product group differ in different parts of the country?

In this analysis, the 30 best-selling product groups were also used, for the reasons already mentioned. In the customer data set six different parts of the country are mentioned, that are: centre, north-west, north-east, south, insular and other. The analysis was performed similarly as in the previous case, so that for each product group the number of events in each part of the country was counted. Given that the number of sold policies varies for different products, but also for different parts of the country, normalization was done by dividing the counted number of events by the number of sold policies for that product, in that part of the country.

Results and Discussion

In this part of the paper, the previously asked questions were answered, by presenting the results obtained by the described methodology in the form of visualizations.

In which months and days of the week do events happen most often?

Fig.1 show the percentage of policy events that took place in each month. The highest number of events happened in July $\sim$ 11.5% and October $\sim$ 10.5%, and the lowest in April $\sim$ 4.5% and March $\sim$ 4.5%.

The second part of the question is answered with Fig.3, which clearly shows that events occur more often on weekdays than on weekends. The largest number of events takes place on Monday $\sim$ 17.8%, followed by Friday $\sim$ 17.8% , while on Tuesdays, Wednesdays and Thursdays, slightly fewer events take place $\sim$ 16%. Weekends have significantly fewer events, with $\sim$ 10.5% of events taking place on Saturday and $\sim$ 6.3% on Sunday.

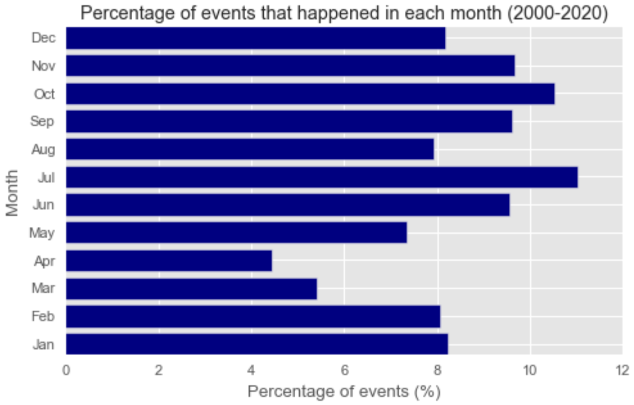

Figure 2. Percentage of events that happened in each month in the period from 2000 to 2020. This figure shows that of the total number of events, most of them happened in July $\sim$ 11.5%, followed by October $\sim$ 10.5%. The month with the fewest events is April $\sim$ 4.5%.

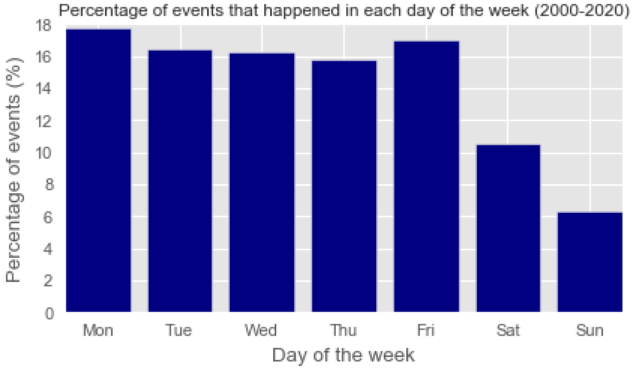

Figure 3. Percentage of events that happened in each day of the week in the period from 2000 to 2020. There is an obvious difference in the frequency of events for weekdays and weekends. Out of the total number of events, the most happened on Monday $\sim$ 17.8%, and the least on Sunday $\sim$ 6.3%

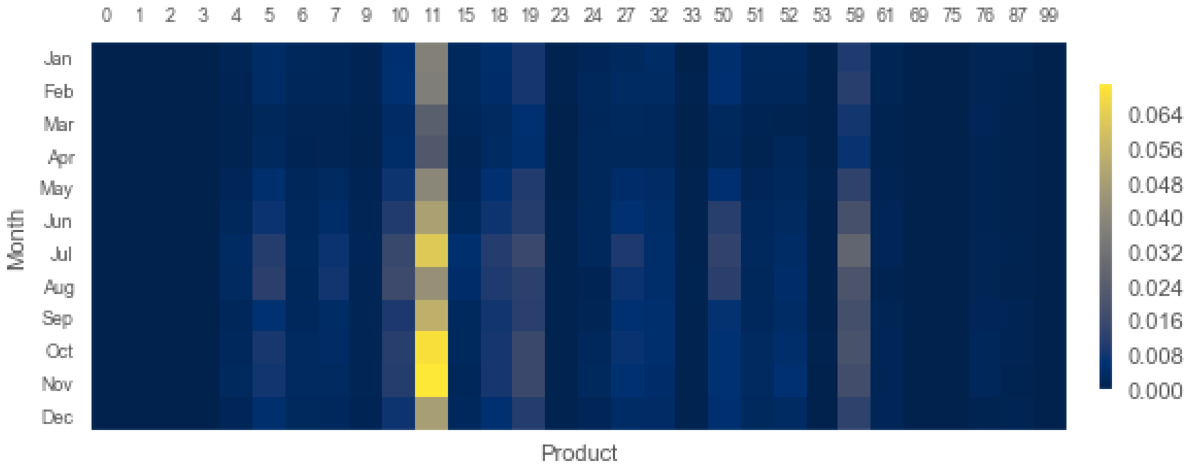

Figure 4. Distribution of the number of events by product groups and months of the year. The number related to each square represents the number of events that occurred for that product in a given month, divided by the total number of underwritten policies for that product. The brightest squares represent the highest frequency of events in relation to the number of policies sold for that product group.

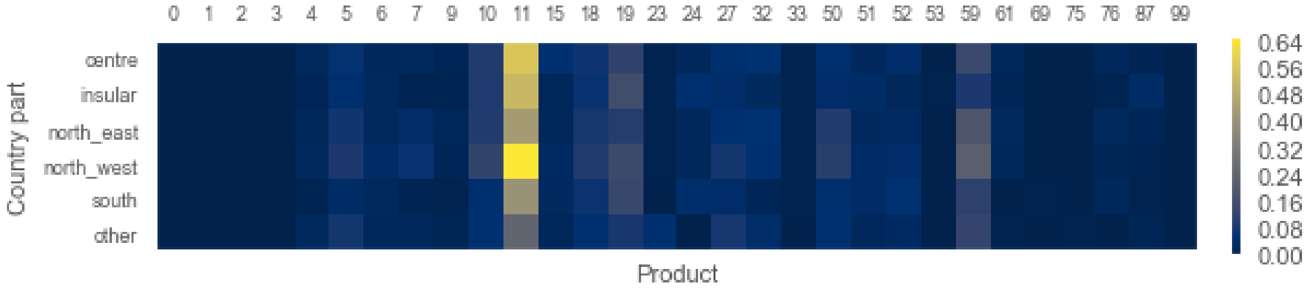

Figure 5. Distribution of the number of events by product groups and different country parts. The number related to each square represents the number of events that occurred for that product in a given country part, divided by the total number of underwritten policies for that product in that country part. The brightest squares represent the highest frequency of events in relation to the number of underwritten policies.

How does month of the year affect the number of events happening for different product groups?

Fig. 4 shows that product group 11 has the most claims in relation to the number of underwritten policies, followed by products 59 and 19. The highest number of events for product 11 occurs in October, November and June. For the majority of the 30 best-selling product groups, the most events occur in July, which coincides with the results shown in Fig.2.

Fig. 4 also shows an unexpected phenomenon, which is that some of the best-selling product groups have no claims at all, such as 0, 1, 2, 3, 33, 53, 69, 75 and 99. This was unexpected because some of them have over 300 000 underwritten policies (product groups 1 and 2).

Does the number of events per product group differ in different parts of the country?

The number of claims per product differs in different parts of the country, as shown in Fig.5. The most noticeable differences are for the product 11, which has the highest frequency of events in relation to the number of policies. As many as 64 claims per 100 underwritten policies occur for this product in the north-western region, which sounds quite unlikely. However, since it is not known what kind of product it is, it is not possible to discuss whether such deviations are expected or some irregularities are present.

Overall, the north-west country part has the highest frequency of claims for the vast majority of analyzed products. The exceptions are product 15 which has the highest frequency of events in the central part, products 24, 51 and 87 with highest frequency in the insular part, product 52 with highest frequency in the southern part and product 76 witch has the most events in the north-eastern country part.

Conclusion

This project analyzed patterns in the occurrence of events that require the payment of claims. Knowing these patterns enables the company to predict future costs, but also to spot potential irregularities that may stem from fraud.