A Machine learning approach in the analysis of Texas consumer complaints on insurance companies

Lenin Kamma

(Class DSC680)

GitHub Portfolio URL: https://databinary.github.io/

Domain

Insurance (Health, Auto, Home, Life, Flood and others).

Data

The data that will be used for this project is going to be extracted from the insurance complaints information system stored and maintained by the Texas Department of Insurance (TDI). As required by law, TDI stores consumer complaints data and facilitate the resolution of complaints. These complaints are submitted by consumers to report unresolved claims or bad practices from insurance agents or agencies. This project will use the data from the years 2018 and 2019. The total number of data records extracted for the year 2019 is 6240 and for 2018 is 12406. The total number of records is 18646.

Data Description

There are 11 features which are of interest for this project. The data description of all these input variables is given below.

Complaint number:

This is a unique number identifying a specific complaint owner:

Received date:

The received date is the date that TDI received the complaint.

Closed date:

The closed date is the date the complaint was closed

Correspondent location:

The correspondent location is the region of Texas where the correspondent is located. These regions are determined by zip code. Zip codes beginning with 75 are referred to as "NE Texas.", 76 are referred to as "N Central Texas", 77 are referred to as "SE Texas", 78 are referred to as "S Texas", and 79 are referred to as "W Texas." All other zip codes are referred to as "Out of State." When the zip code is not available, the record will state "Unavailable.

Subject of complaint:

The subject of complaint is a description of the person or entity against whom the complaint is filed.

Subject name:

The subject name is the name of the person or entity against whom the complaint was filed.

Subject ID:

The subject ID is a unique number assigned to each subject person or entity.

Line of coverage:

The line of coverage is the type of coverage about which the correspondent complains.

Reason for complaint:

The reason for complaint is a description of the reason the correspondent is complaining

Disposition of complaint:

The disposition of complaint describes the resolution of the complaint by TDI

Type of complaint:

The ICIS summaries and download files include confirmed complaints

URL

The data is extracted using the following URL

https://www.tdi.texas.gov/consumer/icis/index.html

References

The following references provide details on how to conduct complaint analysis. These resources give guidance on identifying companies with significant consumer complaints and patterns in the type of complaint. These resources also provide an in-depth view and valuable insights on the present state of insurance companies and the total duration to resolve consumer issues.

[1] **https://www.texasattorneygeneral.gov/consumer-protection/financial-and-insurance-scams/insurance**

TDI accepts written complaints against insurance companies, health maintenance organizations (HMOs), insurance agents or agencies, and other persons or entities regulated by TDI. Complaints generally involve such matters as claims and benefits, false advertising, misrepresentation of policies.

[2] **https://www.usa.gov/consumer-complaints**

This website provides consumers the right to complain about an item or service they purchase.

[3] **https://consumerfed.org/consumer-complaints/**

Consumer Federation of America is a nonprofit research, advocacy and education organization. CFA does not handle consumer complaints but provide guidance and general information on shopping for goods and services

[4] **https://www.naic.org/documents/prod_serv_marketreg_mah_hb.pdf**

The NAIC has developed this Market Analysis Handbook in order to assist states in developing, implementing, and coordinating market analysis programs to regulate insurance markets

[5] **https://www.consumerfinance.gov/data-research/consumer-complaints/**

This database is a collection of complaints about consumer financial products and services that they sent to companies for response.

[6] **https://dfr.oregon.gov/help/Documents/2311-14.pdf**

This report ranks certain insurers by their complaint records, which are based on the number of confirmed consumer complaints closed by the Insurance Division and the amount of premium dollars written by the insurers.

[7] **https://www.iii.org/article/background-on-insurance-fraud**

III is the Insurance Information Institute. Since 1960, it has been the trusted source of unique, data-driven insights on insurance to inform and empower consumers. It serves consumers, media and professionals seeking insurance information.

[8] **http://www.insurance.ca.gov/01-consumers/120-company/03-concmplt/**

The California Department of Insurance (CDI) was created in 1868 as part of a national system of state-based insurance regulation. The insurance market place has changed dramatically over time, but consumer protection continues to be the core of CDI's mission.

[9] **https://www.mckinsey.com/industries/financial-services/our-insights/the-growth-engine-superior-customer-experience-in-insurance#**

This report gives best practices to grow the insurance business

[10] **http://sbp-brims.org/2018/proceedings/papers/latebreaking_papers/LB_7.pdf**

This research proposes a computational approach to characterize the major topics of a large number of online complaints. This approach is based on using topic modeling approach to disclose the latent semantic of 1371 GEICO complaints.

Research and Analysis

There has not been much research done on consumer complaints on insurance companies. As part of this project, I will perform a thorough analysis of the complaints filed by consumers in the state of Texas. The dataset extracted for this project provides many opportunities to perform analysis and predict patterns in the complaints and the zones where the consumers are facing huge problems from the insurance companies.

- What kind of complaints are consumers filing?
- Which insurance companies or agents are targeted in these complaints?
- How long it took to close the complaints?
- Is there any correlation between the state zone and complaint type?
- Can we identify the practices of insurance companies based on the cluster of complaints?
- Can we identify any fraud from consumers based on the complaints?
- Which line of coverage is subjected to the most complaints?
- Can we classify complaints based on the reason for complaint?

Methods & Machine learning models

The input data file has many text fields like the reason for complaint, disposition of the complaint, line of coverage, and subject name. There are two date fields which give the complaint received date and complaint closed date. I will be using both unsupervised learning and supervised learning in this project to know about different patterns in the data, categorization of the complaint, complaint analysis, and prediction of complaint resolution period for different complaint types in the future. The following methods will be used.

- Exploratory Data Analysis

- Scatter plots
- Bubble charts
- Heat maps
- Clustering (K-means)
- Association
- Regression techniques to predict the complaint resolution period

Challenges

- The effectiveness of the clustering method depends on the type of distance selected (Cosine distance, Jaccard distance, or Euclidean distance)
- Data imputation for missing values
- Removing duplicate subject names and complaints
- Merging multiple datasets from multiple years
- Many text features need numeric conversion

Conclusion

The goal of this project is to take data from the Texas insurance complaint database and classify the data points to identify patterns in the data. The underlying structure of the data can be used to identify different clusters and recognize any similarity in consumer complaints. The algorithms can also be used to predict the number of days to resolve a complaint. The patterns are used to identify state zones where consumers are being impacted by insurance company bad practices and detect any fraud in consumer complaints.