What is InsurTech?

- InsurTech refers to the application of emerging technology across the entire insurance value chain in order to address existing problems and uncover new opportunities.

- InsurTech delivers user-oriented or data-driven solutions to the insurance industry including automation of business processes, the development of innovative products, and the exploitation of data for underwriting, risk assessment and claim handling.

InsurTech Examples

- **Mobile devices with apps** automate business processes such as reporting claims, purchasing insurance products, and customer service.

- Health insurance company use **wearable technology** such as smart bracelets to track health measurements and reward healthy behaviour for customers seeking a healthier lifestyle.

- Pay-As-You-Drive auto insurance plans or Usage-based insurance (UBI) use **telematics technology** to allow drivers to have premiums tailored to their driving behaviors.

- Life insurance companies use **facial analytics technology** to produce a life insurance quote by analyzing a selfie photo and estimating the relevant data of the insured (e.g., BMI).
InsurTech Financing for 2011-2022

Data sources: Financial Technology Partners' InsurTech Insights Reports

- P&C
- Life
- Health
- Div.

Funding ($ millions) 0 2500 2750 5000 7500 10000 12500 15000 17500 20000
Number of Transactions 0 50 100 150 200 250 300 350 400
Improving Business Insurance Loss Models by Leveraging InsurTech Innovation

Joint work with Changyue Hu, Panyi Dong, Emiliano Valdez
# Data and Risk Analytics with InsurTech

<table>
<thead>
<tr>
<th>Technology innovation</th>
<th>Insurance type</th>
<th>In-house (traditional) factors</th>
<th>InsurTech factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wearables</td>
<td>Life &amp; health</td>
<td>● Age, gender, marital status</td>
<td>● Blood pressure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Pre-existing medical condition</td>
<td>● Heart rate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Family history</td>
<td>● Glucose level</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Body mass index</td>
<td>● Frequency of exercise</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Tobacco use</td>
<td>● Sleep pattern</td>
</tr>
<tr>
<td>Telematics</td>
<td>Auto</td>
<td>● Age, gender, marital status</td>
<td>● Brakes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Driving history</td>
<td>● Acceleration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Credit rating</td>
<td>● Rotation and turns</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Type of car</td>
<td>● Location, weather condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Business or pleasure</td>
<td>● Distance traveled</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● How much you drive</td>
<td>● Driving attentiveness</td>
</tr>
<tr>
<td>Smart Homes</td>
<td>Home</td>
<td>● Alarm and security system</td>
<td>● Window and door sensors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Age, home structure</td>
<td>● Smart thermostats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Home square footage</td>
<td>● Smart locks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Type of roof</td>
<td>● Smoke detector</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Fire safety and protection</td>
<td>● Water and leak detection</td>
</tr>
</tbody>
</table>
Objective

- The goal is to build a predictive loss model for XYZ Insurer’s BOP line of business by leveraging innovative data sources from Carpe Data, an InsurTech company.

Business Owner’s Policy (BOP)
Loss Data (2010 - 2020)

Data contains policy information and rating factors used in-house about each insured business.

Supplemental data source for business-related features

Features describing insured businesses created by InsurTech Innovation
Industry & University Collaboration: IRisk Lab
Business Owner’s Policy (BOP) Insurance

- Business Owner's Policy (BOP) is a commercial/business insurance policy intended to protect small or medium-sized business owners against potential risks.

- It is available and applicable to variety of industries.

- It combines a various insurance coverages, typically including general liability insurance, commercial property insurance, and business interruption insurance.

- The complexity of underwriting and claims, combined with the low volume and bespoke nature of transactions, pose obstacles to commercial insurance embracing InsurTech.
Insurance Data from XYZ Insurer

- XYZ's **BOP historical loss experience from 2010 to 2020**.
- More than **1,200,000** data entries.

**Policy information:**
- Policy Year
- Earned Exposure
- Coverage Limit
- Exposure Base: **LOI, Annual Gross Sales, Annual Payroll**
- Risk Type: **Apartment, Condo/Office, Contractors, Convenience, Distributor, Fast Food, Motel, Office, Other, Restaurant, Retail, Self-Service**
- Coverage Type: **Building (BG), Business Personal Property (BP), Liability (LIAB)**

**Loss experience and in-house predictive model**
- Observed Loss Cost
- Insurance Company's In-house Model Loss Cost
InsurTech Data from Carpe Data / InsurTech Innovations

Carpe Data provided us with real-time, dynamic information from emerging public data sources shed lights on numerous facets of a business: operations, products, services, physical plant, etc.

• **Business Information**: General operation information about a business.
  - `is_home_business, founded_year, opening_hours, description`

• **Firmographics**: Characteristics to segment prospect business.
  - `business_size, company_type, revenue_range`

• **Risk Characteristics**: Various risk attributes of a business.
  - `commercial_cooking_equipment, raw_seafood_and_alcohol (for a Japanese restaurant)`
InsurTech Data from Carpe Data / InsurTech Innovations

- **Classification**: Categorization of a business.
  - *category, segment, NAICS code*

- **Reviews**: Available reviews for a business.
  - *review content, # likes, star rating, response from owner*

- **Webpage**: Details on a business's webpage.
  - *content, title, url*

- **Group**: Features engineered from collected information.
  - *group 1 - group 13*
InsurTech Data from Carpe Data / InsurTech Innovations

• **Next-generation scores & indexes** A suite of indexes on a 1 – 5 scale targeting dimensions of risk that can be tuned by segment and location.

**Scores:**

• Negative Keywords
• Proximity Combustibles
• Proximity Entertainment
• Proximity Traffic Mode

**Indexes:**

• Customer Rating
• Visibility
• Reputation
• Health & Sanitation
• Maintenance & Condition

InsurTech Data from Carpe Data / InsurTech Innovations

Scores:

• **Negative Keywords**
• **Proximity Combustibles**
• **Proximity Entertainment**
• **Proximity Traffic Mode**

Indexs:

• **Customer Rating**
• **Visibility**
• **Reputation**
• **Health & Sanitation**
• **Maintenance & Condition**
Distribution of Observed Loss Cost

Observed loss cost:

- **Imbalance**
- **Heavy tail**
- **Differ across coverages**
Modeling - Light Gradient Boosting Machine (LGBM)

- A **gradient boosting** framework that uses **tree-based** learning algorithms.

**Advantages** of Light GBM

- Faster training speed and higher efficiency.
- Lower memory usage.
- Support parallel, distributed, and GPU learning.
- Capable of handling large-scale data.
Hyper Parameter Tuning - Bayesian Optimization

- **Bayesian optimization** can effectively narrow the hyperparameter space.
  - Use previous evaluation results to choose the next optimal hyperparameters to evaluate.

- **Distributed learning** by **optuna**
  - Multiple batch jobs on the same model with different sets of parameters.
  - Multiple parameter sets are being trained simultaneously.
Double Lift Charts

- Model lift refers to the ability to **differentiate between low and high lost policyholders** and can be used to measure a model's economic worth.

- Double lift charts are commonly used to measure the model lift and **compare the predictiveness between two different models**.

- Double lift charts are created as follows:
  - Sort data by a ratio of new model prediction (Insurtech-enhanced model prediction) to the current premium (insurance in-house model prediction).
  - Subdivide sorted data into quantiles with equal exposure (we use 30 quantiles).
  - For each quantile, calculate the average observed loss, the average current premium (insurance in-house model prediction), and the average new model predicted loss (Insurtech-enhanced model prediction).

- The model that gives **better predictions** is the one whose predicted loss line is **closer to the observed** one.
Double Lift Charts

Building LGBM MAE Model
Double Lift Charts

Business Personal Property LGBM MAE Model

Double-Lift Chart of Values for BP Train

Double-Lift Chart of Values for BP Test
Double Lift Charts

Liability LGBM MAE Model

Double-Lift Chart of Values for LIAB Train

Double-Lift Chart of Values for LIAB Test
# Model Performance based on Validation Measures

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Dataset</th>
<th>Model</th>
<th>Gini</th>
<th>PE</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>train</td>
<td>Insurance in-house model</td>
<td>0.29</td>
<td>-0.40</td>
<td>5761.94</td>
<td>1526.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tweedie GLM + elastic net</td>
<td>0.44</td>
<td>-0.04</td>
<td>5660.01</td>
<td>1286.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LightGBM</td>
<td>0.84</td>
<td>0.00</td>
<td>5364.05</td>
<td>1198.07</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>Insurance in-house model</td>
<td>0.32</td>
<td>-0.54</td>
<td>5328.02</td>
<td>1461.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tweedie GLM + elastic net</td>
<td>0.32</td>
<td>-0.16</td>
<td>5284.90</td>
<td>1238.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LightGBM</td>
<td>0.37</td>
<td>-0.08</td>
<td>5198.57</td>
<td>1181.47</td>
</tr>
<tr>
<td>BP</td>
<td>train</td>
<td>Insurance in-house model</td>
<td>0.59</td>
<td>-0.07</td>
<td>2498.13</td>
<td>277.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tweedie GLM + elastic net</td>
<td>0.68</td>
<td>0.00</td>
<td>2450.82</td>
<td>262.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LightGBM</td>
<td>0.78</td>
<td>0.00</td>
<td>2409.88</td>
<td>259.11</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>Insurance in-house model</td>
<td>0.58</td>
<td>-0.11</td>
<td>2350.80</td>
<td>270.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tweedie GLM + elastic net</td>
<td>0.36</td>
<td>-0.04</td>
<td>2375.10</td>
<td>262.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LightGBM</td>
<td>0.59</td>
<td>-0.06</td>
<td>2348.93</td>
<td>262.78</td>
</tr>
<tr>
<td>LIAB</td>
<td>train</td>
<td>Insurance in-house model</td>
<td>0.57</td>
<td>-0.67</td>
<td>3937.22</td>
<td>586.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tweedie GLM + elastic net</td>
<td>0.63</td>
<td>-0.04</td>
<td>3920.13</td>
<td>449.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LightGBM</td>
<td>0.78</td>
<td>0.00</td>
<td>3853.67</td>
<td>435.65</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>Insurance in-house model</td>
<td>0.54</td>
<td>-1.15</td>
<td>3347.60</td>
<td>547.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tweedie GLM + elastic net</td>
<td>0.47</td>
<td>-0.33</td>
<td>3340.86</td>
<td>408.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LightGBM</td>
<td>0.56</td>
<td>-0.26</td>
<td>3305.85</td>
<td>394.56</td>
</tr>
</tbody>
</table>
Feature importance:

- Three distinct methods for assessing feature importance.
- With the exception of coverage information, all other significant variables originate from InsurTech.
Illustrative Individual Cases

- To further examine how the InsurTech risk factors affect the loss model, we extracted and analyzed four real businesses from a microscopic point of view.
- Four businesses analyzed are described as follows:
  - (a) a business with a positive claim from the training dataset;
  - (b) a business with no claim from the training dataset;
  - (c) a business with a positive claim from the test dataset;
  - (d) a business with no claim from the test dataset.
A-Land Trust Company B-Rental Apartment

(a) Top 20 influential features of Business A

(b) Top 20 influential features of Business B
(c) Top 20 influential features of Business C

(d) Top 20 influential features of Business D
Implementation Perspective - Residual Modeling

- **Ratio Residual** = Observed Loss Cost / Insurance Company's In-house Model Loss Cost

- Residual modeling has the advantage of improving the predictions *without creating a completely new model.*

- New predictions, however, will be *heavily influenced by old predictions.*
Future Steps

- Insurance tailored feature engineering
  - NLP techniques on text data from social media
  - Spatial and temporal modeling on foot traffic data
- Other modeling approaches
Con(11,9),(990,981)

Concluding Remarks

- InsurTech helps enhance loss model predictions using their databases, otherwise inaccessible by insurers, to gain better insights into the underlying risks.
- This project aims to investigate how much improvement can be gained from these resourceful data.
- Our results indicate substantive differences in the loss cost predictions using real-life data from an insurer’s portfolio of BOP policies.
- This work is an example of the benefits that can be gained from a successful industry and university collaboration through the Illinois IRiskLab.
Selected References


- Steiner, K. and Meng, B. (2019) Predictiveness vs. Interpretability
Selected References


NLP-Powered Repository and Search Engine for Academic Papers

A Case Study on Cyber Risk Literature with CyLit

Joint work with Changyue Hu, and Linfeng Zhang
SOA Report

https://www.soa.org/resources/research-reports/2023/cylit-nlp-search/

https://cylit.math.illinois.edu/
Motivation

The need for advanced literature repository

- No centralized repository of cyber risk literature
  - Limited coverage, e.g., Web of Science & Scopus, Martin-Martin et al. (2018)
- Lack of contextual awareness tool for finding cyber risk literature
  - Keyword-based search, e.g., Google scholar, Beel and Gipp (2009)
- Insufficient integration of the trends in research
  - Static nature and manual review processes of survey papers, e.g., Eling (2020)
Our Solution and Contribution

Leveraging Natural Language Processing (NLP) techniques to automate literature retrieval, summarization, and classification.

- A comprehensive framework with website interface - living literature database and tailored academic search engines
- Application of NLP techniques to improve efficiency and effectiveness
- Applicability demonstrated in the cyber risk domain
Overview

System Architecture

Data Collection Unit

Data Collection
Data Preprocessing
Data Cleaning

NLP Unit
Keyword Extraction
Keyword Clustering
Semantic Search
Association Analysis

Web Server Unit
Web Server
Users
Potential Upgrades using NLP

University of Nebraska–Lincoln Global Research Rankings of Actuarial Science and Risk Management & Insurance™ released by the College of Business at Nebraska.

- Research trend
- Facilitate collaboration
- Dynamic ranking system: innovation and creativity, interdisciplinary connections, public and practical impact, etc.
- ...

Thank you! Q&A