IoT in the Insurance Industry: Using Telematics Data to Strategically Manage Risks and Price Competitively

Simon Jones, Ho-Min Liu, Himanshu Premchandani, Miao Wang, Matthew A. Lanham

Purdue University Krannert School of Management

jone1107@purdue.edu; liu2560@purdue.edu; hpremcha@purdue.edu; wang1894@purdue.edu; lanhamm@purdue.edu

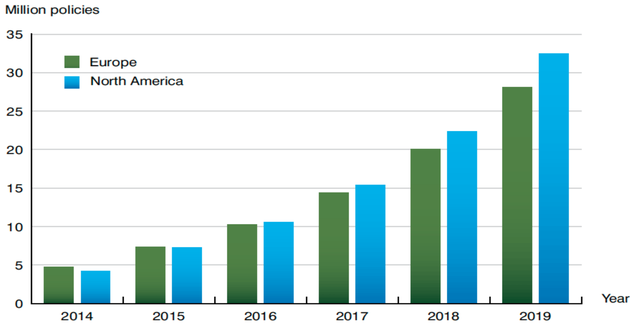
Abstract

The onset of the Internet of Things(IoT) has enabled insurance companies to collect an increasing amount of trip related data using telematic devices that can provide them detailed risk analysis associated with each driver. Our solution estimates the additional particularity telematics data can achieve about insured drivers that could help insurance companies to develop a Behavior-centric Insurance Pricing model.

Our research investigated the differences between the driving behavior of the drivers who have and have not filed the claims. We investigated the presence of differences between driving habits who had claims to see if there is any variability between trips associated with claims and those without to see if certain relationships existed (e.g. if a driver traveling at certain speeds during night hours have a higher risk than those who typically travel with same speed at some other duration of day). We linked the claims data to individual trips data and explored micro factors that might lead directly to the accident cause. Through our research, we also tried to analyze how customers could be incentivized for adopting safer driving habits.

Introduction

- Traditionally, business models for insurance companies are built on complex actuarial calculations and theoretical assumptions
- Revenue of Property and Casualty insurance sector was $558.2 billion last year
- In 2016, average expenditure for auto insurance for an adult in the U.S. was $935.80 dollars
- Telematic devices collect information like location, duration of trip and speed, that can help insurance companies analyze driving behaviors of drivers for developing more efficient premium policies



Insurance telematics policies in force (Europe and North America 2014–2019)

Figure 1. Source: Berg Insight Market Research

Figure 1 shows an increasing trend in demand for insurance policies making use of telematics data. We investigated how we can leverage telematics data into pre-existing actuarial pricing model and in what way it can improve insurance premium pricing.

Literature Review

Good drivers pay less: A study of usage-based vehicle insurance models Authors: Yiyang Bian, Chen Yangb, J. Leon Zhao, Liang Liang	• Traditional insurance and actuarial science has estimated individuals' driving risk based on driver-related information. • Usage Based Insurance (UBI) allows an insurance company to accurately target discounts at careful drivers and charge more on aggressive customers • This research study covers how to utilize massive behavior data to offer assistance for making personalized UBI pricing strategy
The Use of Context-Sensitive Insurance Telematics Data in Auto Insurance Rate Making Authors: Yu-Luen Ma, Xiaoyu Zhu, Xianbiao Hu, Yi-Chang Chiu	• The Brookings Institution suggests that UBI insurance programs incentivize people to cut down unnecessary driving which could potentially yield $50-$60 billion in terms of social benefits. • Hard brakes, hard starts, peak time travel, speeding as well as driving at a speed significantly different from traffic flow are highly correlated with accident rate.
Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data Authors: Mercedes Ayuso, Montserrat Guillen, Jens Perch Nielsen	• As policyholders tend not to be very precise when reporting their average annual mileage, attempts to introduce mileage in the traditional models have not been successful • Some authors conclude that no gender discrimination is necessary if telematics provides enough information about driving habits • The percentage of kilometers per year over the speed limit, the percentage of urban kilometers per year and the total number of kilometers per year show a direct relationship with the number of claims reported to the insurance company.

Data

The dataset we investigated consisted of three separate levels of information:

1. **Drivers -** Contained basic information pertaining to each driver such as birth date, gender, and marital status.
2. **Trips -** Contained trip related information including driver ID, date of trip, start time and end time, distance covered, speed ranges during trip, and other driving behaviors such as speeding acceleration events, braking events, and how much time on phone.
3. **Claims -** Included all the claim associated details such as driver ID, date of claim, incurred amount, and cause of claim.

Methodology

We conducted our research from two different approaches to understand how the given telematics data could be used to get additional insights:

A. Trip-wise Approach: For this approach, we made use of trips and claims data, with a basic underlying assumption that if a claim was made on a particular date by a driver then all the trips taken by that driver on that day were claim-associated trips. Using claim related trips and non-claim related trips, we were able to identify risk level associated with each trip and then aggregate it back to driver level and then use it as variables for modeling.

Feature Engineering:

- For time-period - We calculated proportions of claim related trips to total number of trips for each hour and we did the same for non-claim related trips. From this we engineered two features; percentage of trips travelled at risky time and percentage of trips traveled at non-risky time.
- For speed and percentage-of-speed-limit - For each of the two types of data this process was applied to, we computed two features, representing High-risk speeds and low-risk speeds and high-risk-speeds relative to speed-limit and low-risk speeds relative to the speed limit, respectively. These features were computed as the percentage of miles travelled at these risky speed strata to the total distance travelled.

With new variables aggregated into driver level, models were generated.

B. Driver-wise Approach: Given data collected from each trip from each driver, we were able to inspect, by aggregating data to driver level, whether the distinctions of driving behaviors or habits among drivers would result in different accident rate.

Sample Selection

First entry of Trips' data was recorded on 2017/8/20 (Trip Start). In order to make the duration after the trip start symmetric to the duration before the trip start, we defined sample period as the timeframe beginning on 2016/03/30 and ending on 2019/01/10 so that we have 508 days before and after 2017/8/20. We selected our sample based on this sample period.

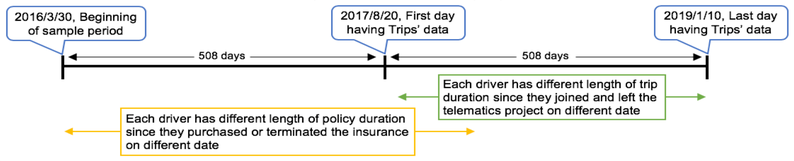


Figure 2. Sample Data

1. Driver Data:
   - Containing basic information pertaining to each driver
   - Focused only on driver whose trips data were available
   - Screened out drivers who had no valid insurance policy in sample period
2. Trips Data:
   - Consisting of start time, date,, distance and speed travelled
   - Removed logically abnormal entries (e.g. average speed > max speed)
   - Aggregated trips to driver-wise average behaviors
3. Claims Data:
   - Containing claim date, driver ID and amount of claim payment.
   - Extracted claims that occurred in sample period (2016/3/30 – 2019/1/10)

After data pre-processing, we conducted following processes:

- Split the sample into 80% of training data and 20% of testing data to evaluate the generalization of the model.
- Since final sample was highly unbalanced, we over-sampled the training dataset using Synthetic Minority Over-Sampling Technique (SMOTE).
- Using trips data attributes as indicators for analyzing driving behaviors we combined them with traditional factors to build prediction models that can foretell the risk level for each driver.
- We used different data mining methods (Logistics Regression, Support Vector Machine, and Decision Forest) and compared the results for final model selection.

We found that Logistic Regression had similar predictive capabilities as the Decision Trees. As logistic regression results are more explainable and computationally cheaper in implementation, we proceeded with logistic regression results to make our recommendations.

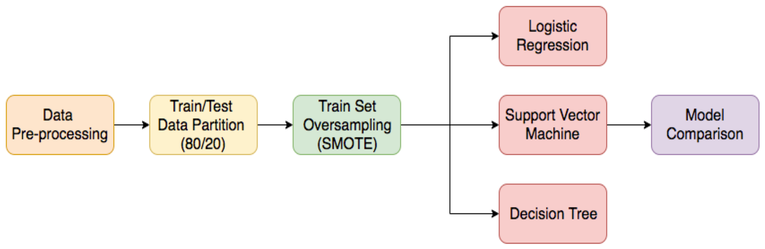


Figure 3. Process Flow

Model Design

The problem we were facing was a classification problem in which we wanted to find out how telematics data can aid to determine prediction on a claim. We first constructed a logistic regression model as:

$$\frac{p_{claim}}{1-p_{claim}} = e^{\beta_0+\beta_1 GENDER+\beta_2 AGE+\beta_3 MARRIAGE+\boldsymbol{\varphi X}} \cdots (1)$$

$$ln(LossAmount) = \beta_0 + \beta_1 GENDER + \beta_2 AGE + \beta_3 MARRIAGE + \boldsymbol{\varphi X} \cdots (2)$$

where p_{claim} is the probability of filing the claim and $log(LossAmount)$ is the natural log form of incurred loss amount; β_1, β_2 , and β_3 are coefficients for traditional variables gender, age, and marital status, respectively; **X** is a variable vector that contains telematics variables such as trip information and driving behaviors; and $\boldsymbol{\varphi}$ are the coefficients for these variables. We also built a linear regression model to investigate the relationship between incurred loss amount and these features.

Results

A. Trip-wise Analysis

- Sample: 276 claim-associated trips and 797,241 non-risky trip.
- At a naïve 50% threshold, model's **Sensitivity** (percentage of claim-associated trips that are identified) is 70% (193/276) and **Specificity** (percentage of true non-risky trips that are identified) is 68% (537,930/797,241).

For the trips-wise categorical analysis, we found decent lift between trips that resulted in a claim and those that did not. In Figure 4 we compared predicted probability of having claim for the class of Actual No Claim and that for the class of Actual Had Claim. Due to the vastly imbalanced data between the two categories, this difference only manifests itself in a change between quartile ranges and means. The remaining differences between these two populations can be attributed to other factors such as risky trips that did not end up leading to a claim, non-risky trips that happened to be flagged as resulting in a claim, general variability among driving patterns and locations not included in our model as well as some random nature of claims. Plotting our regression against our classification (Figure 5), we can see an upward trend of more risk leading to greater expected loss, with frequent notable outliers of low risk-high expected loss and high risk-low expected.

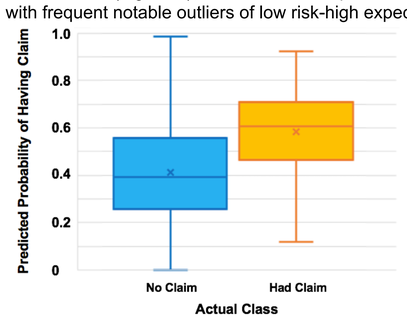


Figure 4. Prediction Comparison

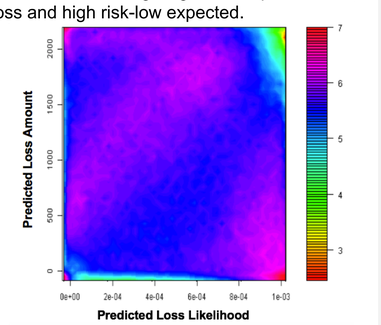


Figure 5. Predicted Features Comparison

B. Driver-wise Analysis

- Sample: 57 claimants (drivers that had accidents and filed the claim) and 1,927 non-claimants (drivers that had no accident not filed the claim).
- Braking events, speeding events, and using phone while driving are highly related with risk.
- New model's **Overall Accuracy** (OA) is 74% while traditional model's OA is 57%.
- New model and traditional model have similar **Sensitivity** (percentage of true claimants that are identified), which is around 60%. (Figure 6)
- **Specificity** (percentage of true non-claimants that are identified) for new model is 74% and that for traditional model is 56%. (Figure 7)

We have found that improvement in accuracy rate results mainly from accurate prediction on the class of Non-Claimants. New model vastly improved by classifying 339 more true non-claimants to accurate class than traditional model. From the perspective of setting premium rates, our analysis suggests the accuracy rate for classifying low-risk drivers using telematics data is better than high-risk drivers. From the traditional model, the company would accidentally raise 339 (840 - 501), which is 17% of the overall sample, more drivers' premiums than the new model. Using the telematics model to set premiums might set premiums too high, which in turn might lead to customers churning to other companies for auto insurance as a result of receiving a rise in premium.

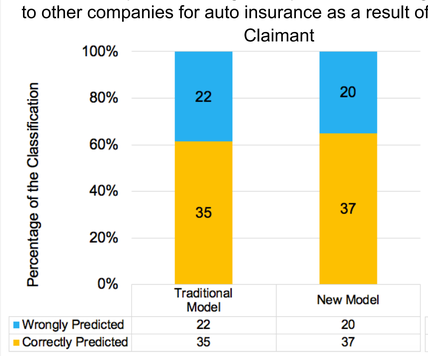


Figure 6. Classification for Claimants

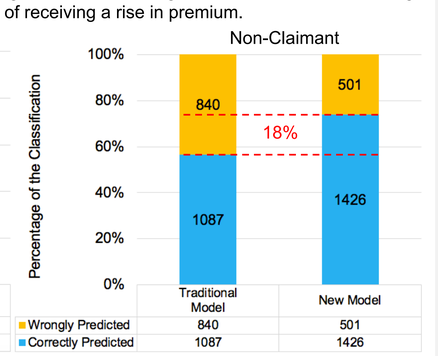


Figure 7. Classification for Non-Claimants

Conclusions

The auto insurance industry has been going through some major changes in recent years with the advent of the Internet of Things (IoT), such as telematics devices recording customer driving behavior. Insurance companies may be able to strategically incorporate this information to provide more competitive vehicle insurance premiums to less risky drivers that deserve a premium reduction. One limitation we found in this study that could be improved in the future if the data is recorded better is with the matching of claim to correct trip. We said that claims were associated to all trips that happen for the given driver on a given day, which means that if a driver takes three trips on average, then ~66% of our flagged claim-trips are not actually claim-associated. Moving forward, if the trip-wise analysis is to be implemented, it would be highly recommended that the company also record the time of the incident, so as to improve the matching process. From driver-wise analysis, the results of our study suggest that the improvement in the accuracy rate of a telematics model (a model that includes driving behavior compared to one that does not) mainly results in better classification of low-risk drivers to the low-risk bracket, compared to identifying high-risk drivers. This is not so surprising as the telematics model includes drivers that have volunteered or been self-selected to participate in the driving program. Those participating in the program will tend to file less claims and drive more cautiously. Incorporating driving behavior in our model thus helps us to more accurately identify and estimate the risk of safer-driving customers.

Acknowledgement: We thank Professor Matthew Lanham for constant guidance on this project.