# Master’s Research Proposal

## *Process Mining to aid in the Predictive Modeling Process*

## Author: Corey Arnouts

# Introduction

There is a field known as Process Mining which uses event logs from Information Systems to look at how Processes play out from a data perspective. Using this event log data organizations can discover how their processes are actually playing out and can work to further improve and optimize these processes. Predictive Modeling is also a field in recent years that has gained a lot of attention. Predictive Modeling uses data from the past to predict what will happen in the future it pairs data with statistical models in order to do this.

Predictive Modeling has a rich history in the Auto Insurance space, using loss forecasting to predict vehicle damages has been a part of the industry for a long time. To a large degree Predictive Modeling in the Insurance space is based off static modeling elements such as the loss history of an Insured, the age of an Insured, type of vehicle, size of home, and Insurance Score which is a metric developed off of credit reporting elements in order to predict future losses.

My proposal is to engineer predictive features that are not only based off of a single point in time but rather are also based on the interactions we have with the client in the past, this will lead to a more dynamic and informative view of our customers. I also think this ground work will also be useful once telematics becomes more mainstream in auto insurance rating because we can start to mine patterns of driving behavior as well to predict claims.

Not only do I want to engineer Features using Process Mining, but I also believe that I can engineer new target variables using Process Mining. Sometimes the outcome variable or business action that you want to predict does not map one to one with a data element that exists within the Information System. In order to be able to predict the outcome variable some transformation and manipulation has to do be done so that the variable is representative of whatever you are trying to predict and Process Mining can aid in more accurately capturing the target variable.

# Objectives

To prove that Process Mining can be used on Complex Event Logs to help engineer both Predictive features and target variables to be used in Predictive Modeling Initiatives, that will ultimately increase the profitability of the company.

# Methodology

### Extract the Data

The first step will be to find all of the needed data from across the organization, this data may be found in different systems and will give different perspectives on how different Business Processes and Client Interactions are playing out. This data will include data that is generated from our employees, our agents, our clients, and our systems. All of the data sources will need to contain timestamps that so they can be assembled in a chronological order that can then be mined for patterns.

### Transform the Data

I will need to put the data in a format in which it can be combined some of the key aspects of Process Mining data are:

##### Case ID: Represents the Party that is going through the process (Person, Policy, Claim, etc..)

##### Activity: The event that has taken place

##### Activity Instance ID: Unique Identifier for the Activity

##### Resource: Person or system responsible for executing the event

##### Start Datetime: start of the event

##### End Datetime: End of the event

All of these data components are necessary to construct a process mining log. Logs of various systems will be combined into one Master Event Log. Once this master event log is constructed data from the target variable will be paired with this event data.

### Analysis

Once the event log data is paired with the target variable the transitions between the events that take place will be encoded with the percent of those cases where the target variable is 1. So for example in the diagram below 11% of claims that travel from the contact insured activity to the first notice of loss activity end up in litigation. All of these transition probabilities are combined and a mean is taken of all of them to determine a Process based Proclivity metric, this methodology was used in the case study below and will be used for the Customer Attrition/Churn case study as well.

After the process mining has generated this feature, the feature will be paired with other relevant information such as policy, claim, account, and geographical attributes among others to improve the improve the potential predictive power of the dataset.

Once all of the Features have been assembled the data will be imported in R, in order to try out a number of predictive algorithms. Various Generalized Linear Models and tree based models such as Random Forest, and XGBoost will be tried on the dataset in order to predict the target variable. Cross Validation on the hyperparameters and the different modeling techniques will be applied to find the most overall predictive model.

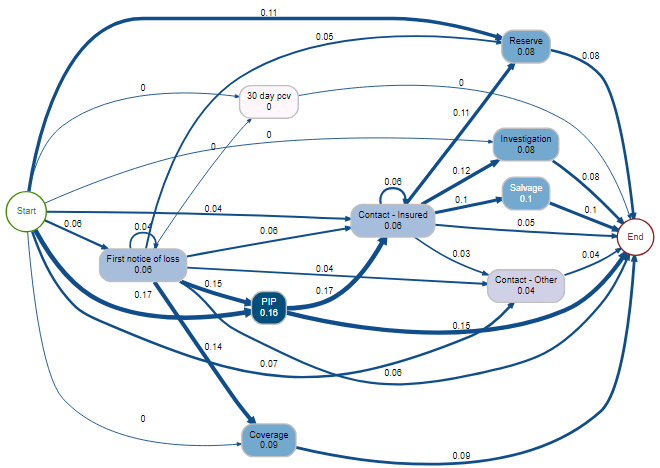
## Evaluation

The evaluation of the predictive model will be done on the test or hold-out set of data, the process based Proclivity Metrics will be applied to the hold out data according to the process patterns that exist with the Hold out data, but this will not cause any target variable leakage because the Process Information that the hold out data will encoded with will only be based off the metrics that were devised using the training data. The different metrics used to compare the predictive models will be ROC, AUC curves and also a gains chart, to see how the well the model is predicting Claims that will end up in Litigation.

# Proof of Concept Use Case:

# Using Event Log data to Predict Litigation on Claims

In Personal Auto Insurance there exists what is known as Personal Injury Protection Claims (PIP) that occur when an insured is injured in a car accident and because of this they need some sort of medical care. Depending on their healthcare plan we may have to act as either the Primary or Excess insurer on this healthcare coverage. Michigan up in till recently was a no fault state and had unlimited limits on PIP Insurance, this has caused an environment of litigation to occur between insurance companies and medical providers essentially because medical providers want to charge as much as possible on injuries that occurred during automobile accidents because they know the Auto Insurer will cover whatever they charge. Because of this PIP Claims are heavily litigated in Michigan and a high percentage of our PIP claims have some kind of litigation occurring on them somewhere in the neighborhood of 10%, litigated claims are increasingly expensive because we then have to pay legal costs and may still be stuck with the full cost of the medical bill, because of this insurers would like to avoid PIP Litigation whenever possible. To do this they would like to identify which PIP Claims have a high likelihood at litigation so they can take actions to prevent litigation.



**First Notice of Loss – The Claim is formally recorded and fully set up in the Information System**

**Investigation – The Claim is under investigation by the Special Investigations Team**

**Contact Insured – The insured contacts the home office**

**Contact Other – Someone other than the Insured contacts the office**

### Process Oriented Metrics (Target Mean Encoding Discrete Event Transitions)

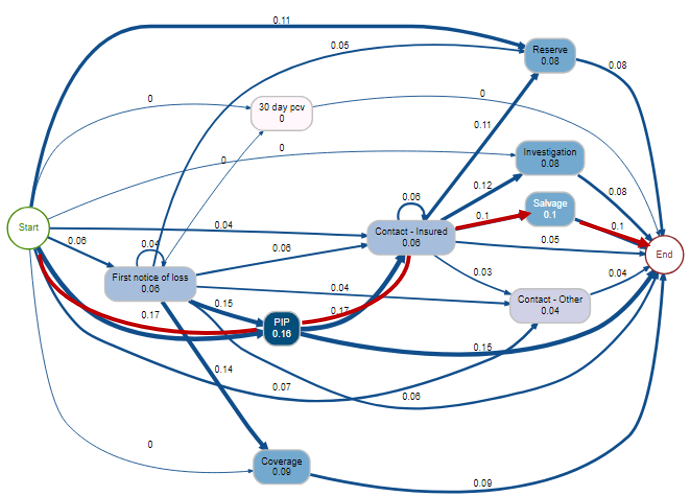
A common practice in Machine Learning is to encode categorical variables with the mean of whatever we are trying to predict, in my case I am trying to predict whether or not a Certain Claim will end up in Litigation, during the Claims Process there is a lot of different steps that take place and in our admin system most of this are indicated

### Process Data

In the Claims Department there is an information system ClaimsCenter that is responsible for handling every aspect of the Claim from the claim being reported to the claim being closed. This information system stores data about the Claim itself, how it was reported, the actions taken on it, the transactions involved with the claim, etc…. Within this system there are Events that are logged on the Claim depending on what happens to the claim, some of these activities are displayed above.

**All of these activities occur during the onset of the Claim (within the first couple months) these activities are only a small subset of all of the possible events that occur on the claim during this time. I paired the Process Event Log Data of historical Claims with a Boolean target variable (0,1) representing whether or not that Claim eventually went to Litigation and looked at the percent of Claims that had a certain sequences of activities and went to Litigation. That is what is represented in the Graphic above. The number within the Activity bubbles represent the number of times that activity occurred on PIP Claims, and the number on the transitions between activities represents the percent of Claims that took that path that also eventually went into Litigation. For example a Claim that has the Investigation activity occur before the First Notice of Loss Activity occurs has a 25% likelihood of going to Litigation much higher than the base rate of 10%.**

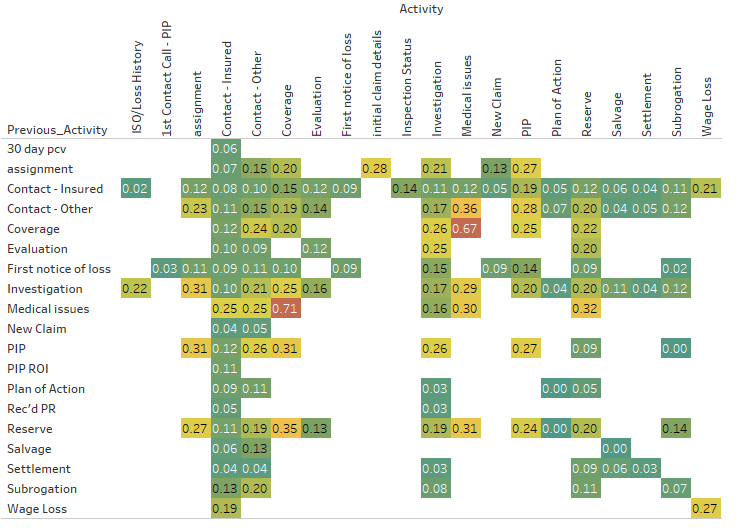
**Coming up with the Process Based Metric (Simplified diagram)**



**The Path above would be represented by the following equation:(.17+.17+.1+.1)/4 = .135, for Claims that followed this specific path their Process based Litigation Encoding would be .135. This roughly equates to saying “based on this Claim’s path only it has a 13.5% chance of going into Litigation”**

## **Engineering Process Based Features**

**Essentially I am using the Activities that occurred on the Beginning of the Claim to understand the likelihood that the Claim will eventually end up in Litigation. I did this on all the Activities Event Log data that occurred on the Claim during the first two months of the Claims Lifespan or before the Claim ever went to Litigation whatever period was shorter. I then came up with an aggregate Metric called Process Based Litigation Proclivity that was the average of all of the Transition Likelihoods. This variable was one of the features that I used in the Litigation Predictive Model. I created a training set of data that was based off of 2017 – 2019 and my test set was based off of first half of 2019 data. The Activity Transition Litigation Likelihoods were mapped to the test set, in order to come up with the aggregate Process Based Litigation Proclivity Metric of these Claims.**



# Literature Review

## Construction of Fraud Prediction Models in Insurance

## Use of Process Mining for Feature Engineering

## Use of Network Analysis for Feature Engineering

## Use of Process Mining for develop a more efficient Claims Handling Process

**[1]** Polato, M., Sperduti, A., Burattin, A., de Leoni, M.: **Time and activity sequence prediction of business process instances**. arXiv preprint arXiv:1602.07566 (2016)

<https://arxiv.org/pdf/1602.07566.pdf>

This paper is probably the most similar of any of the above papers as it is using Transition Probabilities within the event log to predict specific process outcomes which is exactly what I am trying to do. It discusses how there can be difficulties predicting process outcomes when the process is currently running, but this is really the only useful types of models to build. It also discusses certain possibilities like seasonal drift that can occur in process based Predictive Models. To do predictions the models that they build not only use previous traces that look exactly like the process instance being predicted on. They also use Jaccard Similarity and Damerau Levenshtein distance as a way to identify similar Process Instances. This is a really creative way of identifying processes that are similar but are not exactly the same. The different process traces and transitions are engineered into features using one hot encoding. This paper does predictions on the remaining amount of time in the Process Instance and also the future path of the Process Instance. They call the prediction method a Similarity-based Transition system.

**[2] Widad Es Soufi, Esma Yahia, Lionel Roucoules. On the use of Process Mining and Machine Learning to support decision making in systems design.** 13th IFIP International Conference on Product Lifecycle Management (PLM), Jul 2016, Columbia, United States. pp.56-66, ff10.1007/978-3-319-54660-5\_6ff. ffhal-01403073f<https://hal.archives-ouvertes.fr/hal-01403073/document>

This article looks at using both Process Mining and machine learning to optimize the use of resources to achieve a specified Business goal this is what I am trying to do with my project, I want to use process mining and Machine Learning to understand how to better handle PIP Claims so that we incur less costs and bills from Medical Providers.

**[3] Process Mining in Insurance: Measuring and Managing Activity Costs** <https://www.casact.org/community/affiliates/maf/0919/1.pdf>

This is a interesting article because it talks about using Process Mining in insurance to figure out the amount of time that is being spent on various processes within in an Insurance Company a lot of the data elements that they are leveraging are very similar to the ones that I will be using.

**[4] M. de Leoni, et al., A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs, Information Systems (2015),** <http://dx.doi.org/10.1016/j.is.2015.07.003i>

This paper looks at a generic approach to use Process Mining to perform correlation analysis with different target variables and discover underlying patterns. It discusses predicting variables from different perspectives: the control flow perspective, which is predicting the next activity in the process. The data-flow perspective which is predicting a variable about the specific case, this is the one I am especially interested in. The time perspective predicting how much time the specific case will take. The resource perspective, predicting the resource needed to execute the case. This paper really shows how process mining can help predict different aspects of a process, but I would like to extend this to predicting certain characteristics of the actors that are involved in the process, e.g. the customer.

**[5] A Network-Based Approach to Modeling and Predicting Product Coconsideration Relations**

This paper proposes how to use Network Based metrics to improve predictive models. This specific paper discusses how network metrics are used to predict what products that clients will want to buy. It will be helpful to apply a similar methodology when I am evaluating why certain clients are choosing certain medical providers to provide them with service.

**[6] Carlos Andre Reis Pinheiro, Oi, Rio de Janeiro, Brazil. Highlighting Unusual Behavior in Insurance Based on Social Network Analysis**

[**http://support.sas.com/resources/papers/proceedings11/130-2011.pdf**](http://support.sas.com/resources/papers/proceedings11/130-2011.pdf)

This paper discusses how to find unusual connections within the social networks that surround Insurance Claims. I am hoping to implement this in my own predictive model in order to identify which claims could be potentially fraudulent in the future. This paper also uses Principal Component Analysis on all of the Network Metrics to reduce the dimensionality of the dataset. I will also use this kind of methodology to do community detection to see which fraudulent Medical Providers are often working together.

**[7] M. Pospíšil, V. Mates, and T. Hruška, “Process Mining in Manufacturing Company,” in The Fifth International Conference on Information, Process, and Knowledge Management, Nice, France, IARIA, 2013, pp. 143-148, ISBN 978-1-61208-254-7**

This paper identifies ways in which Process Mining can help determine which cases should be escalated to different resources based on the attributes of the Process Instance. This will help construct a methodology for how certain claims should be escalated to more skilled adjusters if the Claim needs to be based off of it’s attributes, both the attributes of the Claim itself and the Claim Process Instance.

**[8] Predicting Insurance Fraud with Machine Learning (SMOTE)**

<https://medium.com/analytics-vidhya/predicting-insurance-fraud-with-machine-learning-smote-da94adf8fb62>

This article takes a look at different techniques to identify future claims, these are some of the variables/features that I will use alongside some of the network analysis and process mining variables. The variables discussed in this article will be some of the same variables that I will plan on using in my analysis on PIP Litigation/Fraud. Some of these variables include client attributes, claim attributes, vehicle attributes, attributes about the injuries that happened during the incident. The article also talks about using SMOTE which is an oversampling technique. This is something that may also employ if I find that I do not have enough positive cases to predict.

# **[9] Artificial Intelligence and Process Mining**

<https://medium.com/datadriveninvestor/artificial-intelligence-in-process-mining-d8a61c0adfd1>

This article really gets to the heart of what I am trying to do with this project, it talks about putting a layer of AI/ML on top of Process Mining to further optimize Business Processes and the outcomes that they are trying to achieve. The Process Mining will generate understanding and new variables that can then be analyzed and help predict the specified target variable. The Process Mining can also inform what the best strategy is to handle the various scenarios that arise within the Claim Process.

**[10]** [**“Process mining on the loan application process of a Dutch Financial Institute. BPI Challenge 2017”**](https://www.win.tue.nl/bpi/lib/exe/fetch.php?media=2017:bpi2017_winner_professional.pdf)Liese Blevi, Lucie Delporte, Julie Robbrecht KPMG Technology Advisory, Bourgetlaan 40, 1130 Brussels, Belgium

This paper is really looking the application of Process Mining to aid in the understanding and improvement of the loan application process within a Dutch Financial Institution a lot of the methodology and measurements that they do can be applied to my analysis of PIP Claims. The better we understand the process the better we can understand how to alter the process to achieve an specified business outcome.

**[11]** [“Predictive Business Process Monitoring with LSTM Neural Networks”](https://arxiv.org/abs/1612.02130) **by [Niek Tax](https://scholar.google.com.au/citations?user=XkRvCC4AAAAJ&hl=en&oi=ao),**[**Ilya Verenich**](https://scholar.google.com.au/citations?user=xRa_fyMAAAAJ&hl=en&oi=ao)**,**[**Marcello La Rosa**](http://www.marcellolarosa.com/)**and**[**Marlon Dumas**](http://kodu.ut.ee/~dumas/)

In this paper they are concerned with Predicting what will happen with running cases of a certain Process Instance. They are trying to predict Process Outcomes based off of what has already occurred int the Process Instance this is very similar to what I am trying to do. I want to predict the result of whether or not litigation will occur based on the previous activity on the Claim. The focus on this paper is to use a Neural Net and try to make this predictions as generalizable as possible so that they perform well on new event logs and new process instances.

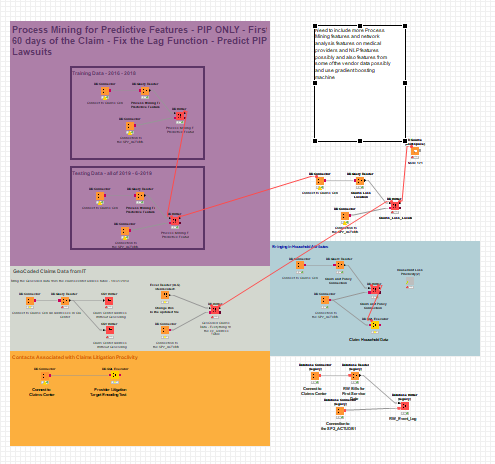
# Analysis and Results

# Data Gathering

For my data gathering I used data from of number of different Enterprise Systems and applications at Farm Bureau Insurance of Michigan. We use Guidewire as our primary administration system and this comes with multiple components.

## Extract Transform and Loading Data

During the Data Gathering phases I had to move a lot of data from disparate sources onto the same server so that the data could be merged, blended, and enriched. In Order to Move Data from Multiple sources I utilized a open source Data Science Platform known as Knime. Knime allows me to execute and an entire workflow of the data science process at once. I spent quite a bit of time in the past connecting Knime to our SQL Servers at Farm Bureau but it has proved to be very beneficial. I can combine SQL Scripts, R Scripts, and Python Scrips all in one cohesive workflow and Knime also has a lot of Data Science Functionality of their own that is accessible.



## Feature Engineering

Feature Engineering is perhaps the most critical aspect of the Machine Learning process and it is where the creativity of person performing the analysis is required.

### Household Metrics

Household Metrics are metrics that are based off of all the business that we have with a specific customer so if they are associated with 10 policies on our book then data from all ten policies will be included. It is somewhat similar to a Customer Lifetime Value metrics that are used in marketing. The Metrics focus on how profitable the Household has been during the time that they have been a client. Households consist of a number of different people who are then connected to a number of policies some of these people may be on the same policy but others may be on different policies but at one time they were on the same policy. Households rules are not necessarily strictly defined at Farm Bureau, but represents connections between people one another and the policies they have or have had in the past. Looking at metrics on this level further let us understand clients on a more holistic basis and client profitability in one line of business has been shown to predict profitability in other lines of business as well.

# 

HHStandardizedRelativeLossCount **=** This metric is derived by looking at all of the policies that the specific household is apart of, and the amount of time that those said policies have been open. This metric is meant to compare how many losses this household has had vs how many losses we would expect them to have based on the amount of time they have been a customer of ours. Then we calculate essientially a Z-Score for the specific policy based off the number of Claims a policy has had vs the number the claims we expected the policy to have. So let’s say on average an auto policy has .1 claims per year with a standard deviation .9. Over a 10 year period we see that a specific policy has 3 Auto Claims so to calculate the Z score of this policy we do the following:

z = (x-μ)/σ), Z= (3 –1)/9, therefore the “Policy Standardized Relative Loss Count” would be 2/9 or .22 which would mean that this policy has a relatively high loss proclivity. Anything over zero is perceived as an above average Loss Proclivity.

We then take Standardized score for every policy that the Houshold has across all lines of business and then aggregate these using an arithmetic average. So if a Household has a Home, Auto, and Farm Policy then we would take the average of the Standardized Loss Counts across all of these policies. This final average metric is the HHStandardizedRelativeLossCount which essentially represents the loss proclivity of an entire household from a frequency perspective.

The reason that the standardization is important is because when you are aggregating across multiple lines of business it is important to understand that different lines have different loss proclivities and different loss distributions so in order to truly calculate a fair metric we have to aggregate Z-Scores as opposed to just dividing by the average. The Z-score always us to account for some lines of business having more variation in their loss frequency distributions.

HHStandardizedRelativePremium **=** Same methodology as above except this features attempts to understand the amount of premium that we are collecting from a client relative to what one would expect based solely on the policies that they have.

HHStandardizedRelativeLossAmount **=** Same methodology as above, but this feature is looking at the loss amount of the Claims, so it is comparting the expected amount of losses for each policy to the actual amount of losses for each policy taking the Z-score based off of this and then aggregating the metric across all lines of Business

HHStandardizedRelativeLossRatio **=** Loss Ratio is a common term that is used in Insurance, it is often used as an Accounting Figure for an entire company or Line of Business it is essientially:

(Losses/Premium Collected), so how much you had to pay for the exposures that you cover vs the amount of premium that you collected to cover those exposures, it is a good indicator of how your company is doing. For this Household Metric I did a similar thing and calculated the Loss Ratios for each Policy and Line of Business and then compared this loss ratios for the expected loss ratio based on the policy type and then aggregated these measures up to a Household Level.

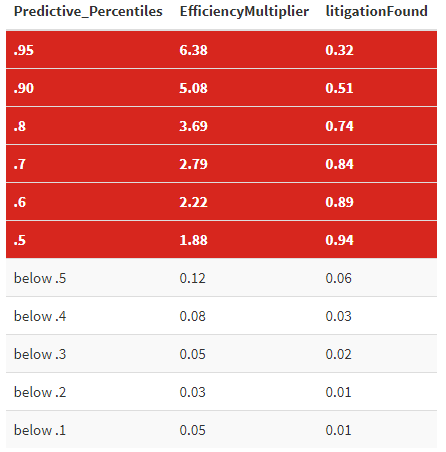
### Claims Contact and HealthCare Provider Metrics

# Evaluating the Model Results

## Metrics that Make Sense for this Initiative

## **Evaluation**

**Using just the Process Based Litigation Proclivities metrics and the geographic location of the Claim the predictive model was able to return a promising results. The chart below shows the results of this model for example the top 5% of highest scoring claims according to this predictive model represented 32% of the Claims that went to Litigation. And the top 50% of highest scoring claims represented 94% of the Claims that went to Litigation.**



**The benefit of being able to identify Claims that are likely to go to Litigation before they actually do is that those Claims can be reassigned to Adjusters that are more equipped to deal with tricky Claims and potentially prevent this Claims from ever going to Litigation or at least limit the amount that we pay out on this Clams. This model allows the insurance company to put the right Claims in the hands of experts. Once implementation time comes I think it may make sense to have multiple models running at once, there may be a model that scores Claims right when they are reported and another model that scores them after the first month, two months, etc… These details still need to be ironed out.**

# **Next Steps:**

## Using Event Log data to Predict Litigation on Claims

For this use case I want to expand on the event log data that I am using, so far I have only used activities and notes from our admin system which are only a subset of the events that occur within the system. I want to see if with this improved event log data I can more accurately predict Claims. I also plan to use network analysis on the parties involved on the Claims to see if I can engineer useful features that way as well.

## Using Network Analysis to Predict Litigation on Claims

The Plan is to also look at the various Medical Providers that are on heavily litigated Personal Injury Claims and determine if there are fraudulent rings that exist, that are looking to purposely overcharge.

## Using Process Mining to Determine the Best Routing for Potentially Litigated Claims

Once I identify which Claims are likely to be Litigated in the Future I would like to determine the optimal path for these Claims to take to avoid as much adverse development on the Claims as possible. This may include certain things like choosing what bills to pay, or what attorneys to hire, or what adjusters should own these potentially harmful Claims. There are also other actions that can be taken on certain Claims things such as Independent Medical Examinations that will inform us whether or not an insured’s injuries are legitimate or if they are being exaggerated.