

# **Using ML To Improve Insurance Policy Coverage**

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science  
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree  
Bachelor of Science, School of Engineering

**Rushil Korphol**

Fall, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Daniel Graham, Assistant Professor, Department of Computer Science

Rosanne Vrugtman Associate Professor, Department of Computer Science

## 1. Abstract

Markel, an insurance based company in Richmond, Virginia required support for their underwriters in order to allow them to create more robust insurance policies at an efficient pace. Working with a team of fellow interns, I created a machine learning pipeline that would recommend forms to attach to insurance policies. This would reduce the time underwriters would spend parsing through thousands of optional forms available for each insurance policy, thereby promoting greater efficiency. Furthermore, forms that would have been overlooked may now be attached, greatly increasing the scope of each policy. The pipeline created was a proof of concept, and much more work is required before it can be commercially used.

## 2. Introduction

Imagine you are given 1 million dollars, but only 1 minute to spend it all. There are an immense number of ways to spend that money, but in such a short time period, deciding what to actually do becomes a tremendously difficult and stressful task. This is the task underwriters are faced with when they must decide what forms to include and exclude to the insurance policies they write. They must carefully pick out the correct forms to ensure that policies are minimized for fraud.

However, insurance fraud is still a major issue in the insurance industry. According to the FBI, over 40 billion dollars are lost to insurance fraud every year. The impact of this is felt by the average insurance holder, who pay between \$400 and \$700 a year from increased premiums.<sup>2</sup> In order to minimize these costs, it is essential that every insurance policy is as robust as possible so that fraud can be lowered. Machine learning is a powerful tool that can be utilized to develop systems that can recommend forms to attach to insurance policies to underwriters, allowing for this cost to be reduced.

## 3. Background

A primary workflow of the insurance industry is underwriters who attach forms to insurance policies to thereafter bind, and formerly issue an insurance policy. During this process, underwriters must make many decisions on which forms to attach and which to omit. The process of underwriters identifying which forms to attach to policies can oftentimes be slow and inefficient during complex cases due to the hundreds of options available. This is especially the case because the policies typically involve large multinational corporations. Certain deals can be lost due to the delays in generating a policy, and even attaching incorrect or unnecessary forms may lead to excess payouts. Furthermore, as stated before, fraud is another major cost

component. Overall, this issue leads to significant losses in revenue.

Machine learning can be utilized to mitigate these costs. Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.<sup>1</sup> Massive amounts of data are fed into algorithms to produce models that serve a wide array of purposes, from predicting the weather to suggesting a new item to buy on Amazon. In this instance, it will be utilized for insurance form recommendation.

There are two primary options for a recommendation system: collaborative filtering and content-based filtering. Collaborative filtering is commonly used in e-commerce scenarios, and works by identifying interactions between users and any items they rate in order to recommend them new items. Content-based filtering identifies features about users' profiles or item descriptions to make recommendations for new content.<sup>4</sup> For example, if a user notes that their favorite color is red, a machine learning system will recognize that and recommend red items more frequently.

## 4. Related Work

There exists an insurance recommendation system that has already been developed and deployed. This recommendation system, created by a group of researchers at the University of Luxembourg, is currently utilized by Foyer Assurances.<sup>3</sup> Their system is for the car insurance industry, and solely focuses on recommending insurance coverages to potential customers. This is a key difference from the goal of my project, which is to recommend forms to policies so that the coverages are improved. However, much can still be gleaned from this work. In particular, it describes the different ways in which the data must be considered when compared to other industries, in that the scale will be naturally be smaller when compared to other cases such as for recommending books or movies. Furthermore, it describes the criticality of ensuring accurate results when compared to those cases.

My project differs from this in that while the previous work aimed to increase the number of people with insurance, I aimed to improve the quality of the insurance coverages. As a result, the data I used was structured differently. Furthermore, the output I provide are recommendation forms to attach to insurance policies, whereas the previous work outputs an insurance coverage. However, there is a clear synergy between both of these systems, and utilizing both at the same time could provide a significant boost in revenue to a business.




## 5. Project Design

## 5.1 Workspace

The entirety of the work was conducted on the Azure Machine Learning platform. Microsoft Azure is a set of cloud based tools developed by Microsoft to allow developers to work directly on the cloud. Due to a lack of experience in machine learning, and the goal of simply providing a proof of concept by the end of the internship, we chose to use this platform. This is because it streamlines the development of machine learning pipelines by providing modules with various functionality that do much of the background work. This would allow us to work at a faster pace and produce a meaningful result. This was particularly useful in our case for generating proofs of concept before moving on to more complex models.

## 5.2 Data

Before a machine learning model can be created, the data must be gathered. Our data was extracted from a database of insurance policy information. The data was structured in a proprietary common policy model(CPM). The purpose of this is to provide a base structure that all the various different types of systems can easily interact with. The information contained consisted of both the insurance policy holder's information such as location, name, size of company, and information about the policy itself, such as what forms were attached. This information was stored in an xml file format. Once the data was received, any private information such as policy holder name and address had to be obfuscated. This is a process whereby all of the strings containing such information had to be distorted so as to not leak anything. After this, the xml files were parsed for the required fields and compiled into two separate datasets, one containing only policy information, and one containing only policy holder information. The policy information dataset was in the form of user-item-rating-triple as seen below. This was a particular format that is required for recommendation algorithms that clearly outlines the relationship between each policy and what form(s) it has. Essentially, the user represents an insurance policy, the item a form that is attached to it, and the rating is simply a 1 to show that it is a valid attachment. These two datasets were then each further split into a test set and a training set. The training set is what a machine learning algorithm will run through to produce a model, while the test set is what the model will produce output from for comparison.

Message ID	Form ID	Rating
 E98D3625-3BF9-43C0-808E-65464F3AE3C8	 MJIL 1000 06 10	 1
E98D3625-3BF9-43C0-808E-65464F3AE3C8	MPIL 1007 01 20	1
E98D3625-3BF9-43C0-808E-65464F3AE3C8	MPIL 1083 04 15	1
E98D3625-3BF9-43C0-808E-65464F3AE3C8	MPML 1003 01 15	1
E98D3625-3BF9-43C0-808E-65464F3AE3C8	MDIL 1001 08 10	1
8FDA9B7A-8BDD-4BFF-	MJIL 1000 06 10	1

Example of policy information data in the User-Item-Rating triple format

## 5.3 Training

After creating the dataset, a training algorithm was then selected. The algorithm runs over the dataset to produce a model, which can then receive new pieces of data and make recommendations from what it has learned. The machine learning algorithm we chose to use was the wide and deep recommender. The wide and deep algorithm is a hybrid recommendation algorithm that combines both collaborative and content based filtering. As a result, it could account for both insurance holder and policy information, providing greater overall results. Furthermore, it was one of the prebuilt modules found in Azure, making it very simple to implement. After the data was fed through the algorithm, a model that could recommend up to 5 forms was created.

## 5.4 Testing

Once the model was trained, its output was tested. The way I chose to measure the performance of the model was to compare the forms that were outputted when the model was predicting for new insurance policies against a list of forms that are allowed for that type of policy. This list was also pulled from Markel's database. I developed a Python script that would make this comparison, and then produce an overall % of valid forms recommended over the whole test dataset.

User	Recommended Item 1	Recommended Item 2	Recommended Item 3	Recommended Item 4	Recommended Item 5
00057602- A899-4883- 99A5- 0B630A289D92 03021878- 1C7C-40A5- A171- 34126C032573 0E21E512- 89EC-4A99- A53F- 39A2EA2F666E 0ED310CF- 92A3-4134- 9D2D- 4FF203678017 188C268C- 5852-4FB5- 89AB- 122134635CED	MPIL 1083 04 15	MEML 5200 02 20	MMX 1207 01 15	MJIL 1000 08 10	MPMX 1000 02 20
	MJIL 1000 06 10	MDIL 1001 08 10	MPMX 1000 02 20	MEML 5200 02 20	MMX 1207 01 15
	MPIL 1083 04 15	MJIL 1000 08 10	MPMX 1000 02 20	MEML 5200 02 20	MMX 1207 01 15
	MJIL 1000 06 10	MPMX 1000 02 20	MDIL 1001 08 10	MPIL 1007 01 20	MML 1004 01 16
	MJIL 1000 06 10	MPMX 1000 02 20	MJIL 1000 08 10	MEML 5200 02 20	MMX 1207 01 15

Example output with 5 forms recommended for each policy

### 5.5 Results and Challenges

The deliverable of the technical project was a machine learning pipeline that could recommend forms to attach to insurance policies for underwriters. A proof of concept was created in Azure Machine Learning that achieved an accuracy of 83%. This meant that of the forms recommended, 83% were a reasonable form to recommend to be attached to a given policy.

An initial difficulty we faced was that there was a slow ramp up as there was a lot of initial confusion on how to begin development as a team and divide up tasks. Although I had prior experience working in an agile team setting through CS 3240, when left on my own I found it hard to act on what I had learned. This ultimately led to a more rushed development of our machine learning pipeline, and as a result it was not as fully fleshed out as it could have been. Another challenge was the lack of understanding of ethical considerations, which meant that additional time had to be spent for self-education, which further slowed down development. This was especially noticeable when obfuscating the data, as that was not a concept I had touched upon during my coursework.

### 6. Conclusion

The final product showcased that it is possible to ethically generate machine learning models without stepping over an individual's privacy as it made use of obfuscated data. It was also able to demonstrate future potential for a larger product that could ultimately be utilized to maximize revenue. My prior experience from taking the Machine Learning course helped me to provide a baseline of context for the rest of my team, which helped to get us started working earlier. Although there was limited growth in my technical skills, I gained a lot in other areas. Specifically in managing and scheduling meetings consistently with a

team, and constantly communicating with team members in order to work together at certain points. One area where I did not feel as prepared was with the obfuscation of data and considering the ethical aspects when creating machine learning models. The issue is that ethics with regards to computing is treated as a one off unit, when it should be woven throughout their curriculum. Multiple scenarios could then be explored with an ethical context, making them more applicable.

### 6. Future Work

There are many limitations of the machine learning model that was created as it was a proof of concept. The data was limited to a subset of types of insurance policies, and therefore is only able to make recommendations for that small set of policies. For further iterations, the data could be expanded to the full range of insurance policies to create a more comprehensive model. Additionally, the model should be trained using a machine learning specific package such as tensorflow or pytorch, which would allow for a much greater functionality. For example, a more robust algorithm could potentially be selected, as there was a limited set of algorithms available on Azure. Additionally, more flexible data transformations would be available, and the program would be far more efficient.

### 6. References

- [1] By: IBM Cloud Education. (n.d.). What is machine learning? IBM. Retrieved October 14, 2021, from <https://www.ibm.com/cloud/learn/machine-learning>.
- [2] FBI. (2010, March 17). Insurance fraud. FBI. Retrieved November 29, 2021, from <https://www.fbi.gov/stats-services/publications/insurance-fraud>.
- [3] Lesage, L., Deaconu, M., Lejay, A., Meira, J. A., Nichil, G., & State, R. (2020). A recommendation system for car insurance. *European Actuarial Journal*, 10(2), 377–398. <https://doi.org/10.1007/s13385-020-00236-z>
- [4] Spetalnick, H. (2019, May 1). Building Recommender Systems with Azure Machine Learning Service. Azure Blog and Updates | Microsoft Azure. Retrieved November 29, 2021, from <https://azure.microsoft.com/en-us/blog/building-recommender-systems-with-azure-machine-learning-service/>.