DSC630

Joel McMillin

Milestone 4 – Finalizing Results

October 29, 2022

Introduction and Proposal Background:

I have spent the last decade working in automobile insurance. A key component of insurance is appropriately rating the level of risk that a customer brings when signing up for a policy, whether it is simply a liability policy or includes first-party property damage or medical coverages. Datapoints frequently used for rating this risk include gender, age, vehicle garaging location, accident history and credit scores, among others. Not every state throughout the country allows the same factors to be used, namely credit scores (DeNicola 2020). As more states restrict which datapoints or demographic factors can be used, it is necessary to find new or alternative ways to rate and price policy premiums.

Throughout the course of this class, I hope to research and provide solutions for the following business problem: How can automobile insurance companies use accident data and insured rates across a given population (that is, what percent of a given population carries insurance versus what percent of that same population could be classified as 'uninsured') in order to better classify risk, and rate and price policy premiums. As part of this, I plan to investigate a couple of datasets that include auto accident details (Moosavi 2022) and insured rates and driving habits (Borelli 2022; Datopian 2012; Timmons 2022). I will also be searching for additional datasets throughout this course that could provide more variables to factor into the models.

Method:

In order to select the best model/models to investigate the business problem outlined above, I want to consider the benefits provided by certain models. I want to make sure that I minimize bias and

variance in the model and recognizing that the data I have available is not continuous data. That is, a simple linear regression might not be the best option. My project is essentially a classification problem, which would be best served by using a model like logistic regression or a decision tree. I want to use a decision tree model because it will be able to use nominal and continuous data, and the datasets I've found so far contain both. Using the decision tree method will help with simplifying the process of using both nominal and continuous data. Another benefit of using the decision tree method is the built-in variable selection ability. They are also a good option when there are many variables being used in the model. The last feature of decision trees that is particularly attractive is that they have the potential to function without manually imputing missing data (Abbott 2014). I have previously had more experience using logistic regression models than decision trees, and so I may opt to use a logistic regression model because they are straight forward. I am also more familiar with interpreting results from logistic regression models than decision trees. A possible drawback is that logistic regression models require a lot more data cleaning than decision trees.

I chose to do a Decision Tree model since my project is a classification problem by nature. I also wanted to use a Random Forest, but I ran into issues with the program running correctly; more information on that later. Using a Decision Tree model, I split my data into training and test sets to fit to the model, following which I completed a couple of evaluation methods. I continue to look for more information on whether any other approaches could be useful prior to final project submission. Using a Decision Tree model and evaluation method were 'successful,' but I would still like to see if any other model might have better results or if I can further fine-tune the variables to get a different, more meaningful result. At this point, the Decision Tree model still seems to be the best available option, which is why I moved forward with that.

Result Evaluation:

I had hoped to use the Area Under the Receiver Operating Characteristic (AUC-ROC) in evaluating my model, but that proved unsuccessful due to errors and limitations I didn't fully understand. However, that evaluation method was one I had in place for evaluating a logistic regression model that I did not ultimately move forward on. As I chose to use a Decision Tree model, the method I used for model evaluation involved using a confusion matrix visual and a Chi-Squared statistic selector to pick the 5 best features in the data and then re-evaluate the model accuracy using those features.

Anticipated Learning:

The goal of this project was initially to determine which factors from automobile loss history and insured vs. uninsured driver percentages could be used to predict loss likelihood (low, moderate, high). With that information, recommendations could then be made to stakeholders about using that information to aid in rating and pricing auto insurance policy premiums. I hoped to learn if the datapoints that I have located are able to demonstrate predictive value with respect to this goal. Ultimately, the goal of the assignment changed to where I am looking for how accident severity can be predicted using a variety of different factors present at the time of a given auto accident.

Proposal Risk Assessment:

This project is not without risks. Thus far, I have been able to identify the following risks based on the project proposal itself and the data I have found:

- The data used is without predictive potential and no greater learning is able to be shared with stakeholders
- Data integrity issues – The data found may be missing something important that I am unaware of/unable to be aware of (are helpful variables missing from the data)
- Data in the datasets is missing or incorrect (are entire entries missing, as opposed to just null or NaN values)

- Have I asked the right question(s)?
- Are any assumptions being made sound, reasonable, correct?
- Do I have enough data to successfully model the data, regardless of its predictive potential?
- Have I selected the correct/best model(s)?

Contingency Plan:

If my proposal as-is cannot be completed for currently unknown reasons, or if the models show no predictive potential, then I had hoped to take a portion of the data to create a separate but related project. By using the loss data by zip-code or city and state, and determining areas with higher loss rates, a customer input tool can be made to help them evaluate their own likelihood for auto losses. This is more simplistic than the outlined project, but will still require model creation and evaluation, and the Python input tool. This can then be presented to stakeholders as an option for customers to get an idea of the type of risk they bring to the insurer and would serve as an early predictor for what their policy premiums may look like. Another contingency plan is for the zip-code data to be used in conjunction with a dataset I found that predicts driver drowsiness (Raju 2020). Stakeholders could be advised that price/premium incentives could be offered to customers in high-risk zip-codes that use the driver-drowsiness webcam insert in their vehicles.

As I have started to work with my datasets, the initial contingency plans outlined above are still attractive, and I have not ruled out using them. However, I found, and will explain in greater detail below in my Preliminary Analysis, that there appears to be a better option that will not require starting 'from scratch,' and requires less rework.

Preliminary Analysis:

As I began to analyze data for my project, I found that I may not be able to answer my question with the data that is both available and relevant to the business problem I chose. I am unable to find

data that shows a breakdown of uninsured drivers as a percentage of the total population by zip code. The data available provides a breakdown of uninsured drivers only at the state level, and unfortunately, this is not sufficient for my initial proposal. Insurance premium pricing is based on far smaller geographic regions than the state or city level, and therefore state-level data will not provide the level of specificity needed to move forward with my question as originally posed. To reiterate, my original question sought to use zip code information in order to help insurance providers to more effectively rate and price their policies. I had hoped to use zip codes, which cover relatively small geographic territories, albeit with poorly defined boundaries in some cases, in conjunction with uninsured population percentages in a particular zip code to better classify risk. The problem of finding appropriate data has required me to rethink my original question and business problem.

After giving the situation some thought, I found an alternative approach that I had not initially considered as part of my contingency plan. Given my setback with finding necessary data, I am going to continue to use the accident data set I have from Moosavi, and I will adjust my driving question. *I now plan to address the question of whether accident severity, as measured (level 1 through 4 rating, where 1 is low severity and 4 is highest severity) in the dataset I found can be predicted using a combination of factors, some environmental and others controllable*. The second piece to this, which still relates closely to my original question and thought process is whether an insurance provider could use this information along with other information gained through telematics to serve as an additional component to rating auto insurance policies. This second piece may, however, be outside the scope of this project. With respect to the other datasets I found, I may still use uninsured driver data at the state level since zip codes do not typically cross state lines. Using some score for this data would then be applied to all losses that occur, by zip code, within a given state. As I decide on which datasets to use for this project, one helpful step will be to continue visualizing the data to get some idea of what the data I have represents.

Whether all my current data is used or not, I did start cleaning and visualizing the larger Moosavi dataset, which contains almost 3 million rows. I have also exploded my categorical variables and am currently working with roughly 140-145 variables, including both the categorical and continuous variables, but I plan to use some type of feature selection after the initial model to see if those can fine-tune the model further. This means the variables will be reduced to the 5 most impactful variables for the final model evaluation. I have spent most of my time cleaning the data and need to continue exploring more useful visualizations than what I have done so far, with most of the visualizations being heatmaps, histograms and scatter plots completed at various stages of data cleaning. The information I have gleaned from this process thus far is just enough to show me that the majority of what I will learn will be through the models themselves. There are no clear relationships within the data, except obvious points such as a day-time accidents being mutually exclusive of night-time accidents. The visualizations also helped me recognize which variables should be continuous versus categorical, with numeric severity needing to be treated as categorical, for example. Another great reminder through this process has been how to treat zip codes. While numeric in appearance, zip codes are effectively categorical, and can be tricky to work with in Machine Learning. For this reason, I have kept the longitude and latitude variables to better pinpoint loss locations in relation to accident severity. At this point in the process, the most helpful visualizations are those that helped me to better understand the scope of the available data. As the project progresses, and now that I have modeled the data, I am going to look for effective ways to visualize my findings.

My intention was originally to use Logistic Regression or a Decision Tree approach to model my data. The Decision Tree option is ultimately what I chose to use, and I believe that it will help in answering the question while reducing the amount of data cleaning needed to accomplish it. The modeling used for the Decision Tree option is also an area I want to gain more experience with due to its popularity. I think this is a situation that warrants the added effort honing newer and less familiar skills,

and using the Decision Tree approach will do just that. My position on model evaluation will continue to use the Chi-Squared metric which will then be used for top feature selection for a final run of the model using those. While my question has changed, my data has not changed, nor has the way in which my data will be used.

While my data hasn't changed, and my model and evaluation approach do not look as though they will need to be changed, my underlying question and business problem have changed from the original proposal submitted, allowing me to still use a Decision Tree model. Throughout the preliminary analysis I have sought to provide clarity to the change in the driving question as well as the ramifications of that change on the other components of this project – the datasets, the models and the evaluation and interpretation of the models. The dataset itself is large enough, both in variables and in sheer volume of automobile accident data, that it is still useful for the purposes of my project. The nature of the data itself and the question being asked are such that no other major changes need to be made at this point. However, my original expectations are not reasonable insomuch as what I had hoped to accomplish will not be feasible, and that's fine. I am confident that my path forward will provide interesting insights into what factors can be used to predict automobile accident severity in such a way that insurers can mitigate risk while consumers can be better educated on what controllable and environmental factors contribute to the likelihood of severe accidents.

Data Preparation:

In order to prepare my data for modeling, I reviewed the variables and made determinations based on my experience working in insurance about which might be relevant for the business question I posed. The dataset, prior to creating any dummy variables, has over 40 variables to pick from. There are obvious ones to remove, such as the instance ID, and others that are not so obvious, such as those that indicate traffic markings or weather conditions. Some of these are relevant in that a consumer could

possibly avoid driving in certain types of weather, while others, such as visibility, as measured in miles, would not have an impact on what is within a driver's span of control. Since I am looking at ways consumers and insurers can avoid high severity accidents, it makes sense to focus on variables that are most within a consumer's control. That was the guiding principle for my initial data cleaning.

My dataset had over 2 million observations, so I was easily able to remove rows with null values without overly reducing the dataset's total observations. I did have instances of negative values (for longitude and latitude, as well as some outliers in other variables which seemed to be coded incorrectly), and for these, I did scale them, but then ran into errors later with trying to complete feature selection using the Chi-Squared metric. Needing positive numbers made sense, but the scaling failing to resolve that was confounding, as the scaling I chose should have provided values between 0 and 1, and not -1 and 1. Due to this issue, I used absolute values of variables to ensure they were all positive. A concern I had with this was with longitude and latitude, using positive instead of negative values would change real locations. However, the goal of this project is to provide recommendations on factors affecting accident severity that insurance companies can look at to better rate policies. With that in mind, insurers being aware that there might be a connection to specific locations 'in general,' suffices insomuch as the findings could be used to emphasize to insurers the need to account for driver location when looking at loss severity. These locations could change significantly over time due to infrastructure changes, among other factors, so the impact of longitude and latitude don't exist in a vacuum: insurance companies might need to visit and revisit location and loss severity repeatedly over the course of time.

Build and evaluate at least one model: See Jupyter Notebook for models

Results Interpretation:

The model I finally used was a Decision Tree, as previously noted. I did also use a Linear Regression as I built up to the Decision Tree, in order to rule out any linear relationships within the data, and I quickly accomplished this as seen in the accompanying code. I also wanted to use a Random Forest model in addition to the Decision Tree, but there were issues with the model fitting where the code simply would not run. I investigated this issue to see if I could find anyone else with a similar error. I followed the recommendations such as reducing the number of instances run, but this did nothing to resolve the issue. Finally, in addition to the Decision Tree model, I also found a method for using the actual text descriptions of each loss, a variable I had previously discarded due to potential for too much noise in that data, to find relationships between the descriptions and accident severity.

The Decision Tree model at first appeared to be fantastic, with model accuracy at 87% initially, and then 89% after optimizing the model. However, when I made predictions for severity based on this model, the predictions were all for Level 2 Severity. Early visualizations showed that there are far more Level 2 Severity losses than there are of any other type. This might indicate the data will inherently overfit due to that data imbalance. Scaling the data didn't make any difference here, but I did want to look at feature selection to see if anything could be learned through that. The first method for feature selection that I used, which used the Chi-Squared scores of variables to select the features, showed that Severity itself, which should not have been a possibility, was the top feature – so it seems the model may have suffered from overfitting in addition to the imbalanced data. The second method I used gave a slightly different set of variables, excluding Severity, which is more in line with what I wanted to see. This second method showed the top 5 features as follows:

```
Variable: Start_Lat            Importance: 0.41
Variable: Blowing Snow Nearby  Importance: 0.36
Variable: Start_Lng            Importance: 0.21
Variable: Humidity(%)          Importance: 0.02
Variable: Day                  Importance: 0.01
```

These variables seem to show that there are in fact variables, not including accident Severity itself, that do have an impact on accident severity. These have potential for serving as the basis for recommendations to insurance companies with respect to preventing high severity accidents and incentivizing safe driving for consumers. At the same time, the final accuracy index measured showed a score of 1.0, which means that the model's accuracy is essentially nil.

The final model that I used that involved the text descriptions was interesting in that what it showed was very interesting, but it essentially confirmed what one would expect: the heavily weighted features involved redundant words and phrases like "accident," "traffic," "closed" – which we would expect to see in relation to any accident regardless of severity. The low weight features also had similar words and phrases. So, while interesting to experiment with, this was not useful for final conclusions or recommendations.

Initial Conclusion and Recommendations:

At this time, based on the results of my Decision Tree model and feature selection, as well as my own experience working in auto insurance, the information found shows interesting, albeit expected results: Location of accidents, as measured by longitude and latitude, might impact accident severity – I say 'might,' since the modeling results ranged from 87-89% but then also showed the distribution is essentially random with the 1.0 index score. Whether the location is a high traffic intersection without traffic lights, or a high-speed highway with closely situated on and off-ramps, certain locations will lend themselves to higher severity accidents. Insurance companies have the opportunity to use their knowledge of these locations to incentivize drivers avoiding these locations and thus higher severity accidents. Weather conditions, like humidity or presence of blowing snow, indicate factors that might not necessarily be easily controlled by a consumer or incentivized by an insurer (for example: How many residents in northern states can avoid driving when there is snow present? How many residents in

southeastern states can avoid driving when there is high humidity or rain present?). Currently, my recommendation would be for insurance companies to use telematics to incentivize drivers avoiding high risk traffic areas, but with the added recommendation that they also focus on ways to work with the automobile industry to increase safety features that could help drivers during inclement weather.

References:

Abbott, Dean. “Chapter 13 - Case Studies.” *Applied Predictive Analytics*, John Wiley & Sons, Inc., Indianapolis, IN, 2014, pp. 214-215.

DeNicola, Louis. “Which States Restrict the Use of Credit Scores in Determining Insurance Rates?” *Experian*, Experian, 21 Apr. 2022, https://www.experian.com/blogs/ask-experian/which-states-prohibit-or-restrict-the-use-of-credit-based-insurance-scores/#:~:text=homeowners%20insurance%20policies.-,California,much%20you%20pay%20in%20premiums.

“Explanation of Receiver Operating Characteristic (ROC) Curves.” *GetTheDiagnosis RSS*, http://getthediagnosis.org/roc.html.

Ghosh, Samadrita. “The Ultimate Guide to Evaluation and Selection of Models in Machine Learning.” *Neptune.ai*, 21 July 2022, https://neptune.ai/blog/the-ultimate-guide-to-evaluation-and-selection-of-models-in-machine-learning.

To-Date Data Sources:

- Borrelli, Lena. “Uninsured Motorist Statistics and Facts 2022.” *Bankrate*, https://www.bankrate.com/insurance/car/uninsured-motorist-statistics/.
- Datopian. “Bad Drivers.” *DataHub*, https://datahub.io/five-thirty-eight/bad-drivers#data.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. Moosavi, Sobhan. “US Accidents (2016 - 2021).” Kaggle, 12 Mar. 2022,

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents., arXiv preprint arXiv:1906.05409 (2019).

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Moosavi, Sobhan. “US Accidents (2016 - 2021).” Kaggle, 12 Mar. 2022, https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents. In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.
- Raju, Serena. “Yawn_eye_dataset_new.” *Kaggle*, 7 July 2020, https://www.kaggle.com/datasets/serenaraju/yawn-eye-dataset-new.
- Timmons, Matt. “Uninsured Motorist Statistics: Changes by State and over Time.” *ValuePenguin*, ValuePenguin, 4 Aug. 2022, https://www.valuepenguin.com/auto-insurance/uninsured-motorist-statistics.