Premium Indication Tool

[redacted]

Western Governors University

A. Proposal Overview

The project seeks to create a model that can predict a potential premium, formally known as an indication, for an insurance customer using features related to customer demographics, recent incident details, and requested policy requirements. The goal is to provide customers with a quick indication of what their premium could be if they choose to continue to an official premium quote. An important distinction is that this will be an indication only, not a binding quote. The dataset being used is specific to auto insurance, but the means and methods could also be modified and made useful for any type of insurance premium indication prediction.

A.1 Research Question or Organizational Need

Using a dataset of 1000 car insurance claims, I would like to build a prototype that would take in a simulated new customer application and indicate potential premium. Included in this project will be data cleaning, feature selection, model evaluation, a new applicant form, and a machine learning model that will predict potential premium indications.

A.2 Context and Background

Most insurance companies rely on underwriters who have proprietary and complicated formulas and processes for analyzing risks. This is often a time intensive process which leaves customers waiting for extended periods of time to find out insurance premiums. What I would like to do is create a workflow that would provide an instant indication or estimation of premium based on a form submission. Although this would be an estimate only, it would give prospective customers some idea of the premium cost. I believe this would be beneficial to insurance companies. The option of an instant estimate would be an advantage over competitors and draw customers to the company and increase brand exposure in the market.

A.3 and A3A Summary of Published Works and Their Relation to the Project

1. "Nationwide and Bold Penguin partner on Agent-Facing Commercial Insurance Quoting and Selling Platform"

This article discusses Nationwide's partnership with Bold Penguin and their goal to create an insurance quoting platform for independent agents (Nationwide Mutual Insurance, 2021). The need is for a quick, easy, and versatile solution that will streamline the process for agents and provide speed and efficiency for customers. The two companies working together have developed this platform and are testing it with a small group of customer agents. This is a strategy that insurance companies of all types are seeking to employ. This is one of many articles that support the validity of the project I am attempting to undertake. My goal is similar in that I would like to build a rating engine that provides quick and easy access to instant quoting for both customers and agents.

Nationwide Mutual Insurance. (2021, September 22). Nationwide newsroom. Nationwide and Bold Penguin partner on Agent-Facing Commercial Insurance Quoting and Selling Platform.

https://news.nationwide.com/nationwide-and-bold-penguin-testing-new-platform/#:~:text=To%20help%20agents%20meet%20evolving,with%20solutions%20from%20multiple%20carriers.

2. "What's The Difference Between a Quote Vs. Indication?"

This article written by Bigfoot Insurance outlines the differences between a bindable quote and an indication. Indication is a non-binding rate based on several answers provided by the potential applicant. (User, 2023) This differs from a quote in that a quote is binding. This means once the quote has been offered and accepted by the customer, the premium amount is official. This is an important distinction to be made in relation to the project. This project does not aim to calculate a bindable quote. The idea is to create an indication that will be used to quickly inform customers and agents of a potential premium amount. A final bindable quote must be created using more strenuous formulas and calculations. Automating a bindable quote is a different project altogether.

User, G. (2023, August 28). Quote vs. indication. Access One80. https://accessone80.com/blog/quote-vs-indication

3. "An improved random forest classifier for multi-class classification"

In this paper, the author's goal is to use and improve a Random Forrest Classifier to make multiclass classifications on disease in crops, specifically groundnut disease (Elsevier, 2016). The article first explains the use of classification. Classification is the recognition of categorical data described by a set of features in the dataset. The author then discusses the idea of feature selection and the methods of testing and selecting the right classification model for the dataset. These are the same steps I will take in this project. Analyzing and cleaning the data, feature selection are the first steps. Then I will run tests against Random Forrest, Neural Network (NN), Logistic Regression (LR) and Support Vector Machine (SVM) to determine accuracy and best fit.

Elsevier, B. V. (2016, September 1). An improved random forest classifier for multi-class classification. Information Processing in Agriculture. https://www.sciencedirect.com/science/article/pii/S2214317316300099

A.4 Summary of Data Analytics Solution

The solution will consist of a statistical test known as Spearman Correlation coeffeciaent. This test will analyze correlations between features and the target variable. Correlations between features and the target will be accessed during the feature selection process and a random forrest classification model will be used to asses prediction accuracy. The result will be a form that a prospective customer or agent can use to get a quick indication of premium.

A.5 Benefits and Support of Decision-Making Process

The initial benefit will be to prospective customers or agents. This will be a tool that will streamline their process, giving them quick and easy access to the information they need to decide on which coverage they need. As they shop around, the tool will allow them to select different limits and deductible combinations to figure out which product is right for them. This will also reduce processing time for sales and customer service personnel at the insurance company. Customers and agents who use the tool will be armed with the specific coverage they need and already have an idea of the premium they will be paying before they engage sales and customer service personnel.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

The goal of this project is to find correlations between features and premiums in the Claims dataset and use the correlated features to create a classification model for predicting premium indication.

1. Determine if there are correlations between features and premiums.
    a. The deliverable for this objective will be a barchart showing correlations.
2. Build a classification model to make premium indication predictions.
    a. The deliverable for this objective will be a classification model.

B.2 Scope of Project

B.2.A Included in Project Scope

The scope of this project will include a Jupyter Notebook using various Python packages and Python code to collect, clean, and analyze data related to auto insurance claims. Included in the project scope is Exploratory Data Analysis to determine correlations, model training and evaluation including accuracy, precision, recall and F1 metrics. as well as an interactive form for demonstrating real-time prediction.

B.2.B Not included in Project Scope

Building formulas and calculations for bindable quoting is out of scope. The indications the project will provide are simply a tool that can be used to get an idea of what premiums might be given the form inputs.

B.3 Standard Methodology

For this project I will use a Kanban approach or pull system. Each milestone listed below will be sub- grouped as its own project. Each milestone will be completed and tested before moving to the next milestone. Each step will be documented with comments within the project for future reference.

B.4 Timeline and Milestones

Milestone/Deliverable: Acquire data from Kaggle

Duration 1 hour Projected start date: 11/1/2023 Anticipated end date: 11/1/2023

Milestone/Deliverable: Data Cleaning

Duration 1 hour Projected start date: 11/2/2023 Anticipated end date: 11/2/2023

Milestone/Deliverable: Correlation Analysis

Duration 1 hour Projected start date: 11/2/2023 Anticipated end date: 11/2/2023

Milestone/Deliverable: Build classification model(s)

Duration 1 hour Projected start date: 11/3/2023 Anticipated end date: 11/3/2023

Milestone/Deliverable: Build interactive form for demonstration

Duration 1 hour Projected start date: 11/3/2023 Anticipated end date: 11/3/2023

B.5 Resources and Costs

•. Hardware item: No cost, I will be using a fully depreciated Macbook Pro.

•. Facilities: No cost, home office will be used

- Software: No cost, Python and Jupyter Notebooks are both open-source and free.
- 10 work hours: No cost, I will be providing all the labor for this project.

Total Cost: $0

B.6 Criteria for Success

The successful project will consist of correlation analysis determining which features, if any, in the dataset are statistically significant. Features in the dataset with p-values less than .05 will be considered statistically significant.

C. Design of Data Analytics Solution

In this part, you will discuss the design details of your Capstone data analytics solution.

C.1 Hypothesis

Correlations can be made between a policy's annual premium and other features in the dataset including customer demographics, recent incident details, and requested policy requirements.

A machine learning model can be created to predict a premium indication with accuracy above .9.

C.2 and C.2.An Analytical Method

Pandas provides the method 'corrwith' which which can be used to compute Spearman rank correlation coeffecients. I will use this method to draw correlations between each variable in the dataframe and the target 'policy_annual_premium.' This method is an appropriate choice as it is extremely useful in comparing a single target with multiple variables and provides me with a correlation coefficient for each comparison. I will then plot these coefficients to analyze their strength in correlation.

Additionally, I will attempt to use a Random Forrest classifier to make predictions for potential premium indications and assess performance based on Accuracy, Precision, Recall, and F1.

C.3 Tools and Environments

The development environment used will be Jupyter Notebooks. Additionally, I will use Python and the python packages listed below.

- numpy

- pandas
- matplotlib.pyplot
- seaborn
- from sklearn.preprocessing - StandardScaler
- from sklearn.impute -SimpleImputer
- from sklearn.decomposition - PCA
- from sklearn.ensemble - RandomForestClassifier
- from sklearn.model_selection -train_test_split
- from sklearn.metrics - confusion_matrix, accuracy_score, precision_score, recall_score, f1_score, classification_report
- matplotlib.imagematplotlib.pyplot
- matplotlib.cm
- operator
- from sklearn.preprocessing - MinMaxScaler
- from sklearn.preprocessing - OneHotEncoder, StandardScaler
- from sklearn.compose - ColumnTransformer
- from sklearn.pipeline - Pipeline
- ipywidgets
- from IPython.display - display

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

Correlation Coefficient

- My null hypothesis is no correlations can be made between a policies annual premium and customer demographics, recent incident details, and requested policy requirements.
- The planned statistical test is a Spearman rank correlation.
- Correlation coefficient.
- The *alpha* value will be .05

Classification Model

- This is a supervised classification model.
- The algorithm(s) to be used is a Random Forrest Classifier.
- The metric(s) to be used to assess performance are Accuracy, Precision, Recall, and F1.
- If the Accuracy is ≥ .9, the model will be considered successful.

This method is an appropriate choice as it is extremely useful in comparing a single target with multiple variables and provides me with a correlation coefficient for each comparison. I will then plot these coefficients to analyze their strength in correlation.

C.5 Practical Significance

This project will be significant if correlations can be made between the policy annual premiums and the variables in the dataframe. If there are no correlations in the dataframe, it will be difficult to make predictions. However, Random Forrest classifiers are good in this situation as they can make decisions even if many strong correlations are not present. Since most of the data is categorical, this is potentially true of my dataset. This project may also require more data and/or additional features to be a total success.

C.6 Visual Communication

I will use matplotlib to create a couple of bar charts to visualize the correlation coefficients for each variable. One bar chart will compare continuous variables after normalization and another to compare categorical variables after encoding. I may also create a heatmap for additional visualization.

Additionally, I will use matplotlib to visualize the decisions made by the trees in my Random Forrest Classifier to analyze feature importance.

D. Description of Dataset

D.1 Source of Data

Sources: Shah, B. (2018, August 20). Auto insurance claims data. Kaggle. https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data

D.2 Appropriateness of Dataset

The dataset consists of 1000 claims including demographic data, recent claim history, and selections associated with the policy. These features can be used to draw correlations between each feature and the policyholder’s current premium. Also included in the dataset is the premium for each policyholder. This is appropriate data for the analysis since it represents the data of current and actual features associated with annual premiums.

D.3 Data Collection Methods

The data was collected by downloading the .csv file from www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data This data was originally gathered from databrinks.com and assembled for data analysis projects such as this one.

D.4 Observations on Quality and Completeness of Data

The dataset looks mostly clean and complete except for one null column. Additionally, all data looks to be relevant to the given columns. However, the dataset is also highly categorical, but the target variable is continuous. This will pose a bit of a challenge when drawing correlations. Steps will need to be taken to make premium indication a class that can potentially be predicted. There are also many columns that are irrelevant to predicting premiums. I will either remove them using strict domain knowledge or correlation analysis.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

The data source was acquired from a public and free repository on Kaggle.com. That source will be and has been documented throughout this project plan and credit given to its owner.

This dataset provides privacy and security concerns. Although there is no personal identifiable information such as names and social security numbers, it does contain demographic data linked to a policy number. Because that policy number can be used to gain more sensitive PII, appropriate action should be taken to make sure it is safeguarded.

An ethical concern to consider is potentially biased or discriminatory decisions made by the model. I do not believe that will be an issue as this project only aims to provide an indication, not the actual premium.

I do not believe the project infringes on any HIPAA violation as it will not return any insurance data back to the potential customer or agent

References

Data: Shah, B. (2018, August 20). Auto insurance claims data. Kaggle. https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data

Purdue University. (n.d.). Retrieved from APA Formatting and Style Guide (7th Edition) - Purdue OWL: https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/index.html

Scribbr. (2022, December 21). *Free Citation Generator*. Retrieved from Scribbr: https://www.scribbr.com/citation/generator/

Smith, J. (2023). A Generic Journal Article Example. *Generic Journal*, 50-62.