Premium Indication Tool

[redacted]

Western Governors University

A. Project Highlights

- 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.
- 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. 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.
- For this project I will use a Kanban approach or pull system. Each milestone 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. Project Execution

Below is the original execution plan for the project. In addition to the barchart showing correlations, I decided to also create a heatmap. I did this just to see correlations between features that are not the target feature. The project timeline was shifted slightly, but still completed without adding additional resources.

The project plan

The goal of this project is to determine if correlations between features and premiums in the Claims dataset exist 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.

The project planning 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.

Project 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

C. Data Collection Process

The data selection and collection process were exactly as expected. 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.

C.1 Advantages and Limitations of Data Set

- An advantage to this dataset is that it is highly relevant and a good representation of the features that are typically associated with insurance claims and premiums. Additionally, the dataset was quite clean to start and well labeled.
- One big disadvantage that I found is most data is categorical which by itself is not a terrible thing. However, the dataset would benefit from more data.

D. Data Extraction and Preparation

To start the data extraction process, I first downloaded theclaims file from Kaggle.com. Before any analysis could be done, I had to also import the file into a dataframe using the pandas library.
Processing tasks

- *Feature _39 was removed as it contained only nulls.*
- *Insured_Zip was also removed as the feature contained too much variability to be useful. The dataset itself is only 1000 records, and contained 995 unique zip codes. With this much variability, correlations would not be made.*
- *Many columns in this datasource were related to the claim outcome. However, the only outcome that I am interested in is Policy Annual Premium. Therefore, I removed other columns that are only outcomes and left columns impacting premium.*

- *Additional checks for null and duplicates were performed, but none were noted. These data cleaning steps are appropriate as they help maintain data quality and consistency for accurate analysis and proper model performance.*
- *All numerical features were normalized using MinMax Scaling. Minmax scaling is needed in order to make sure all features are equally weighted. When analyzed for correlation and placed in the classification model.*
- *All categorical features were encode using Pandas get_dummies. It was important to encode these variable to convert categorical features to numeric/binary features. This is an appropraite choice to make the data more consistent and interpretable.*

E. Data Analysis Process

E.1 Data Analysis Methods

For this project, I used two different data analysis methods, a spearman rank correlation to get a correlation coefficient and a Random Forrest classifier to build a classification model. The project's goal is to identify features from within the dataset closely correlated with policy annual premium. Finding the correlation coefficient using a Spearman Rank Correlation 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. Also, the spearman rank correlation method and the Random Forrest Classifier work well on datasets that are not linear. For this test, a result of 1 indicates a perfect positive relationship. Alternatively, a result of –1 indicates a perfectly negative relationship. A result of 0 indicated no relationship. The significance of each result can then be evaluated using a p-value typically set to .05.

Random Forest classification is a bit more robust. The process constructs a series of decision trees, and each tree votes on which class to move the data point into. The class that collects the most votes is determined to be the correct class. Random Forests also provides a process for determining feature importance which determines which features are most closely correlated to the target feature. This model will be especially useful in making the needed predictions since the data is highly categorical. The metric(s) to be used to assess performance are Accuracy, Precision, Recall, and F1.

E.2 Advantages and Limitations of Tools and Techniques

Because my dataset is highly categorical, Spearman Rank works well. A major advantage is that it works well with no normally distributed data. A disadvantage is that it my sometimes require larger datasets to find relationships between features.

Random Forrest is a perfect solution for this project as it can easily handle non-linear relationships. It is highly accurate when dealing with categorical data. However, it can also be difficult to interpret since they can become quite complex.

E.3 Application of Analytical Methods

I used the corrwith() method to draw correlations between each variable in the dataframe and the target ‘policy_annual_premium. This function requires the pandas package to be imported into the project but is extremely easy to use. Once my data was cleaned, normalized, and encoded, I called the corrwith() function from the claims dataset and passed in policy_annual_premium as my target feature. This outputs a list of all features as well as their correlation coefficient. I then graphed each of those coefficients for visual analysis.

Additionally, I was able to use a Random Forrest classifier to make predictions for potential premium indications and assess performance based on Accuracy, Precision, Recall, and F1. I first defined the categorical and numerical features. Then I created a pipeline for both categorical features to be encoded and numerical features to be normalized. Next, I created a pipeline that includes preprocessing and the classification model. The classification model was defined and limited to 8 nodes. After the pipeline was created, I split the dataset into testing and training data and fit the data to the model. I was then able to sort and plot each feature on a barchart in order of feature importance. The last steps of this process were to split the data into

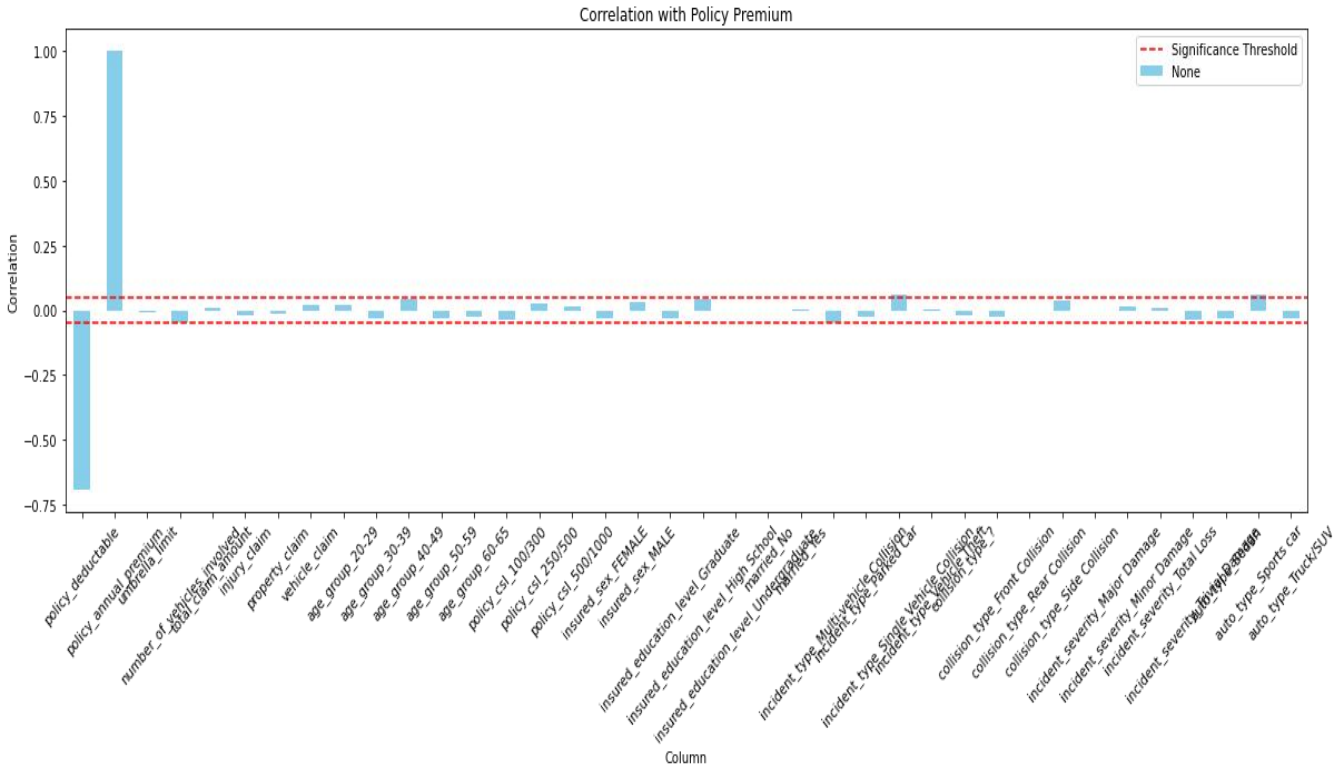
testing (20%) and training (80%) data and calculate Accuracy, Precision, Recall, and F1 using the testing data.

F Data Analysis Results

F.1 Statistical Significance

Statistical significance was analyzed first by calculating a correlation coefficient of each feature to the target variable. In my analysis I noted that policy deductible shows a strong negative correlation to policy annual premium. Significance was access using a p-value of .05. The chart below depicts this correlation very well. For this reason, we fail to reject the null hypothesis. In fact, correlations can be made between a policies annual premium and requested policy requirements. However, my goal is to use this data to predict a potential premium indication. I had hoped to find better correlations to support that goal.

However, random forest classifiers can still perform quite well even given the low amount of correlation found in the dataset. The model was trained with a max depth of 8 and produced predictions with .91 accuracy.



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
- There is sufficient evidence to fail to reject the null hypothesis and support the claim that no correlations can be made between a policy annual premium and customer demographics, recent incident details, and requested policy requirements.

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.
- Although strong correlations could not be made, the random forest classifier does a decent job making predictions as evidenced by these scores.

  Accuracy: 0.91

  Precision: 0.90

  Recall: 0.91

  F1-Score: 0.90

F.2 Practical Significance

The correlation between policy deductible and premium by itself is not significant, but it is still a strong correlation that will play a part in the random forest classifier. With that said, the random forest classifier did an excellent job creating an ensemble of decision trees that are complex enough to make predictions with 91% accuracy. This will provide a level of confidence when creating premium indications. Again, I am not trying to create a bindable or official quote, but rather an indication of what premium might look like given the subset of features selected.

F.3 Overall Success

Based on the results presented in F1 and F2, discuss how the project was successful. This section may repeat content from sections F1, and F2 above, and Task 2 section B6.

G. Conclusion

G.1 Summary of Conclusions

The goal of this project was to find correlations between various features and a policy annual premium and build a model to predict potential premium indications. Finding correlations was not as successful as I hoped. I discovered a strong negative correlation between policy deductible and policy annual premium. This correlation by itself is not significant enough to predict a

premium indication. However, the addition of a random forest classifier was able determine feature importance and better handle non-linear relationships. This resulted in a model that can predict premiums with 91% accuracy. This should be sufficient given the objective is premium indications, not bindable quotes.

G.2 Effective Storytelling

The graphical representation of correlations between the various features and policy annual premium did an excellent job highlighting the strong correlation between policy deductible and policy annual premium. Since the data was mostly categorical plotting data points on an X/Y axis would have been inappropriate. Instead, I used a bar chart to highlight the correlations.

G.3 Recommended Courses of Action

My first recommendation would be to gather additional data. Although the model predicts premium indications, the model would certainly benefit from additional data. Continuing to add current data will keep the model relevant and accurate. My second recommendation would be to continue exploring features that may play a more crucial role in predicting premiums. It is possible that the organization can pull additional features from their database or begin collecting additional data from their customers that will improve the model.

H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=[redacted]

References

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

pydata. (n.d.). *Pandas.dataframe.corrwith#*. pandas.DataFrame.corrwith - pandas 2.1.3 documentation. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corrwith.html

SKlearn. (n.d.). *Sklearn.ensemble.randomforestclassifier*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html