Final Proposal

Daehee Hwang

BU Faculty of Computing & Data Sciences

DX699 O2 AI for Leaders

Dr. Joshua Von Korff

Sunday 20th, 2025

Project Statement

Throughout the course of the class, I employed various statistical and data science techniques to verify the plausibility of my dataset *Insurance_claims* (Aqqad,2023) in identifying fraud in an automotive insurance claim. Both univariate and bivariate methods showed promising results and opened up opportunities to explore questions I have not previously considered. I furthered my analysis using more multidimensional statistical and data science techniques like univariate analysis, PCA analysis, and linear regression.

Every year, property damage fraud is estimated to cost the shareholders $40 billion in fraudulent claims. This accounts to $400-700 of each policyholder's premium going into paying out these frauds. (FBI, 2010) I continue my journey into building a full-feature fraud detection model in this project with the wealth of new information obtained from weeks 8 through 12.

Exploratory Data Analysis: Executive Summary

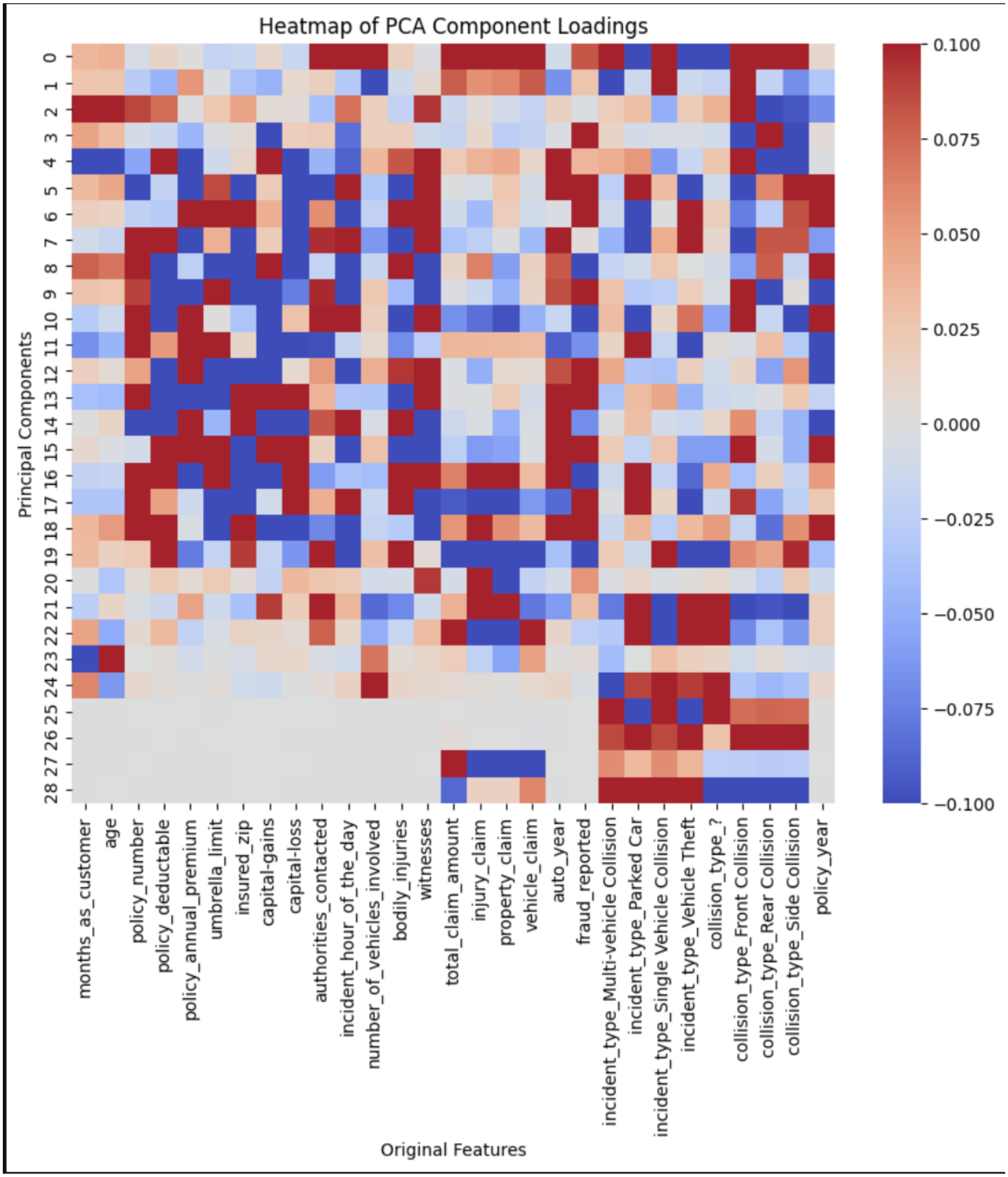
Complex set of data that *Insurance_claims* contained offered many opportunities for analysis. It performed well in both supervised and unsupervised training. Supervised learning offered promising results for identifying relationships between the target variable ('fraud_reported') and other features of the dataset. Unsupervised learning allowed me to identify less obvious patterns and explore its implications on insurance claims. Methods used to run these analyses were principal component analysis and linear regression with both random forest regressor and cross-validation. These were then visualized with bubble plots, heat maps, and chart graphs.

Exploratory Data Analysis: Results

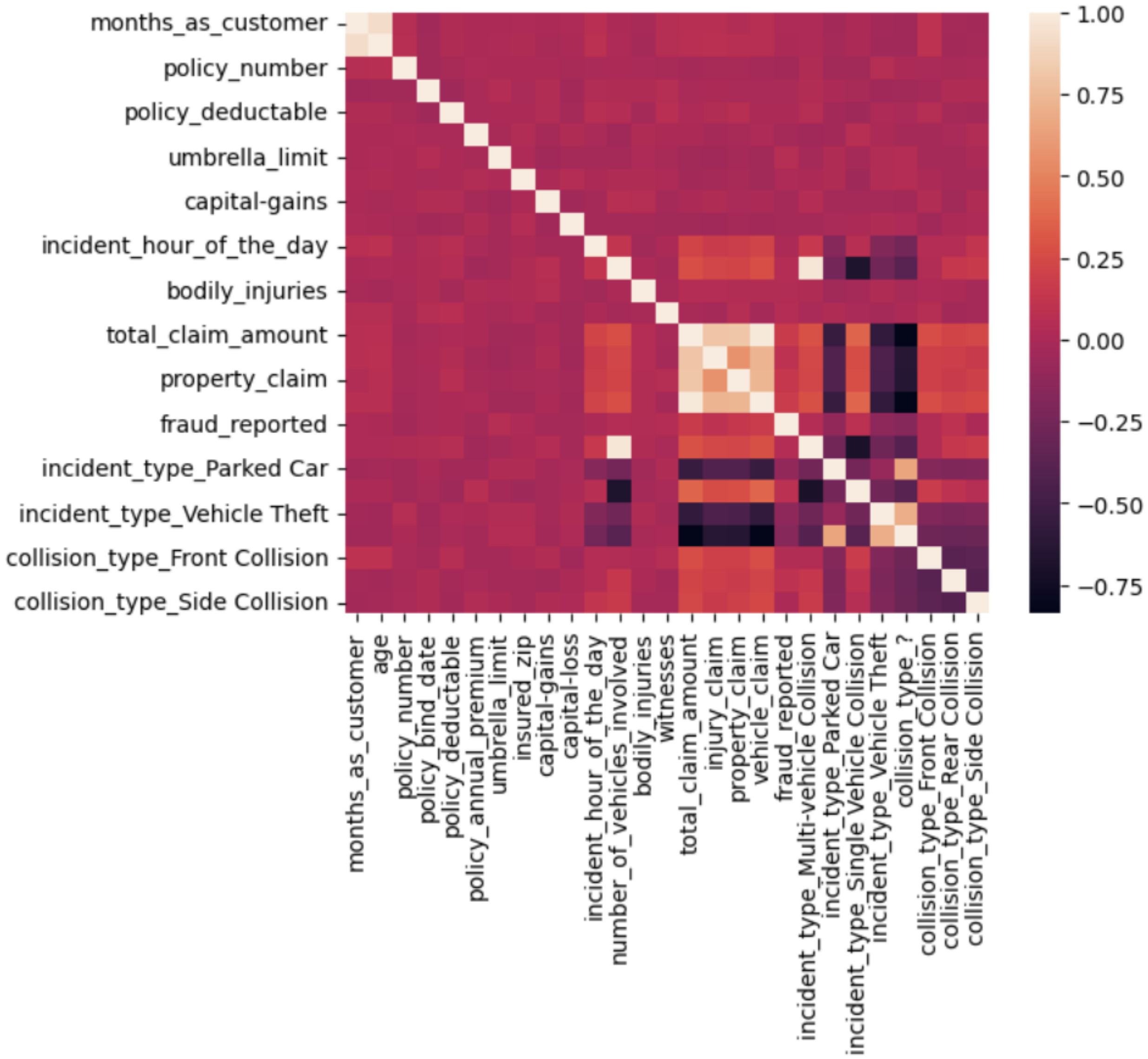
In Milestone Three, the original dataset was cleaned, hot-encoded, and preprocessed to reduce the number of components from the original 40 to 29 which was saved as "processed_insurance_claims." To analyze further opportunities to reduce the dimensionality of the data while maintaining the integrity of the data (variance) I first used principal component analysis. Principal component analysis on the dataset showed that the first component of the data only captured 20% of the variance in the dataset. This may have been due to the large number of features in the dataset. Meaning variance of the dataset is distributed amongst many important features. Even with the first five principal components combined we only capture 64% of the data's variance. Further supporting the hypothesis that the dataset has complex interrelationships that cannot be easily reduced to just one or two dimensions. *Be careful with your principal components (Björklund, 2019)*

While reducing dimensionality in larger sets of components using principal component analysis proved to be a non-viable option, it still offered great insights into which components contributed heavily to the data's variance. There were recurring features in the first five components that consistently contributed high to the variance. These were the financial payout data like we examined in the bivariate analysis in milestone three like total claim amount, injury payout amount, and capital loss. Beyond what was previously identified, the principal component analysis showed that features like incident type and collision type also contributed heavily to the dataset. While the data should not be reduced in terms of principal components, I can use this as a reference when performing further feature selection techniques.

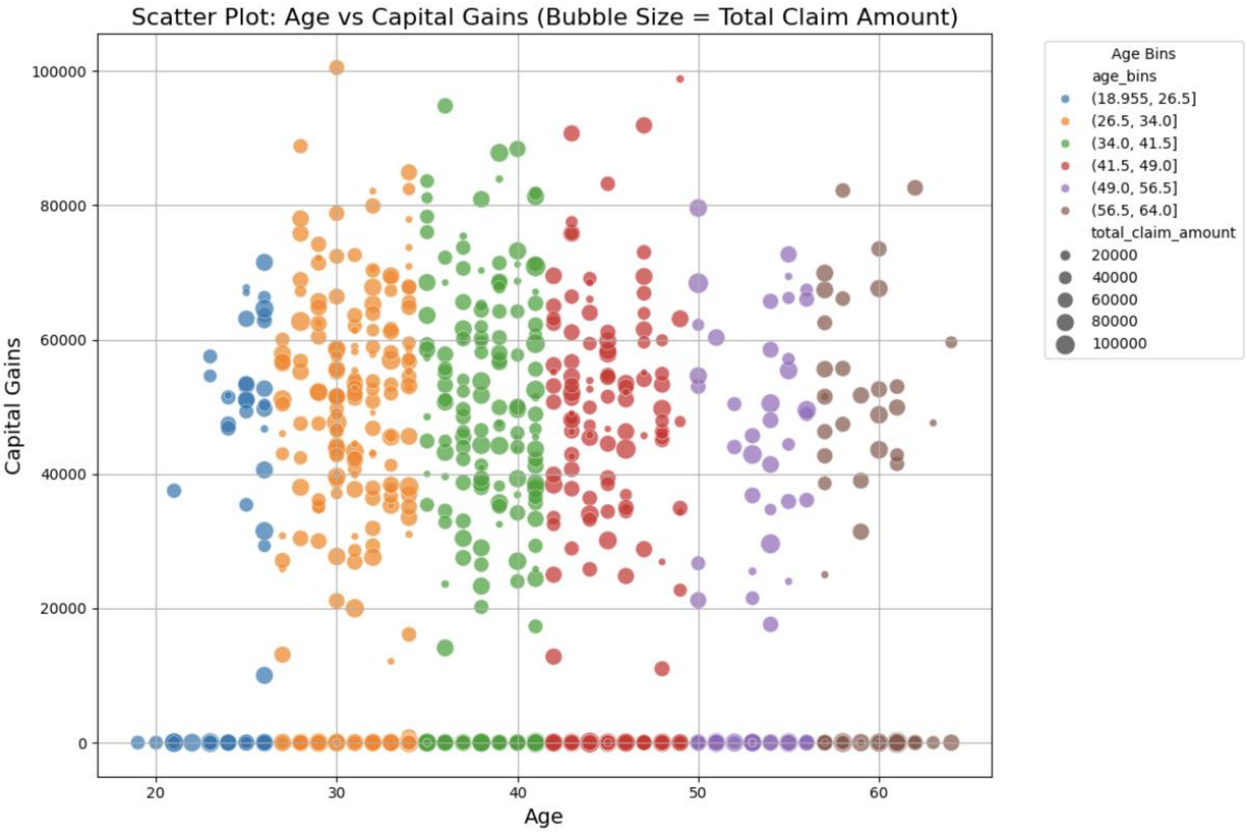
Heatmap of PCA Loadings



Bivariate Correlation Analysis from Milestone



Comparing the feature by feature analysis from milestone 3 to the current analysis, we can see that univariate analysis captured more information about each components' contribution to the dataset. Bivariate correlation analysis could not capture the contribution of collison type and incident type to the data set. Demographical data like age also contributed higher to the overall dataset's variance than previously identified. Using this new information I ran a 3 dimensional visualization using age, capital gains, and total claim amount.

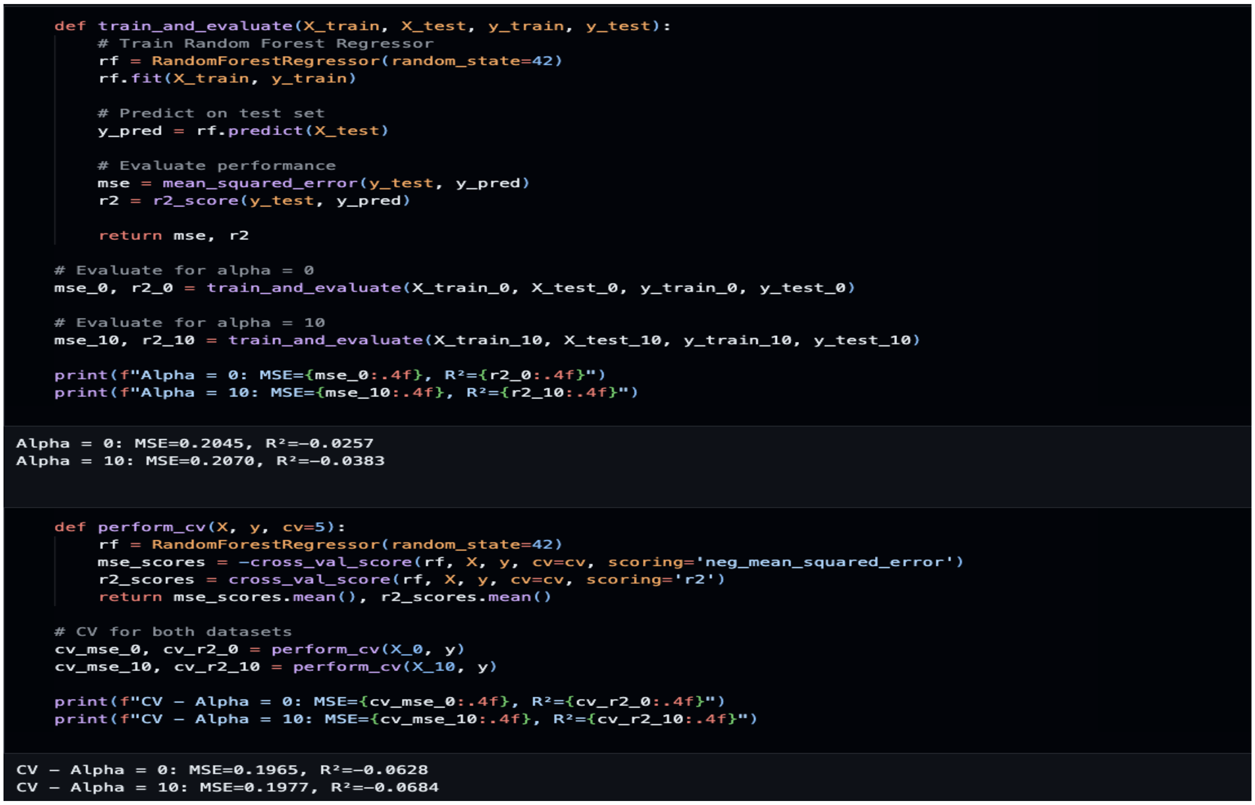


Analyzing the 3 dimensional bubble plot, we can see that while drivers older 56 contributed to a lesser amount of claims - their average total payout on the claim was generally larger than the rest of the age brackets. This may be due to older drivers causing more severe accidents and/or older drivers being more prone to injuries and fatalities *Older Adult Drivers (CDC, 2022)*

I then performed linear regression analysis to further analyze the dependencies and predictability of the components. Basic linear regression using the scikit-learn library function showed that there were obviously more correlated components and some components that could be considered noise to the overall dataset. As we previously identified, financial components(total_claim_amount, injury_claim, property_claim, and vehicle_claim) and incident_type/ collision_type features had r squared value of 1, meaning these components are highly predictable. But it can also mean that these components are redundant of each other and it may be resourceful to reduce these into smaller sets of components. *Rsquared* (Nau, 2019) Demographic information like age and months as customers showed high correlation. Two components that were previously not identified as providing high predictability was the number of vehicles involved and authorities contacted but these had high r squared value. Which may be contributed to the fact that higher number of vehicles involved results in higher total claim amount and large enough accidents where fire department or police have to report will generally mean that the accident was more severe.

Surprisingly, umbrella limit which is the total claim payout limit did not show high r score which was in line with the PCA analysis (first three components had umbrella limit in the low contribution range of 2-3%) and the correlation analysis performed in milestone 3. Other features with low linear regression score was policy number, policy premium, and insured zip. Policy number was in line with it being a random identifier for the claim, while policy premium and zip code is independent to the customer so it makes sense that it does not contribute largely to the dataset largely dealing with accident/incident details. Policy number was previously determined to be statistically insignificant to the dataset but was left during the data processing stage in milestone 3 as I thought we may want to keep the identifier feature in the preliminary analysis stage as a means to identify the specific claim with error or outlier. Now that we are reaching the final stages of analysis. I will be removing the feature from the model. I will be performing deeper feature selection processes to further identify the plausibility of keeping other features identified with PCA and linear regression to be less important to the dataset, like umbrella limit, insured zip, and policy year.

After completion of linear regression, I explored results of random forest regressor, cross validation, and parameter tuning. The random forest regressor showed results in-line with the previous analysis. Limiting the alpha to 10 increased error and lowered the r squared score. MSE was increased to 0.2070 from 0.2045, while r squared score decreased to - 0.0383 from -0.0257. Meaning limiting the nodes of the forest hurt the overall error/confidence rating of the model. This is inline with PCA analysis where I hypothesized that the dataset had complex interdependencies where reducing the dataset in large sets will hurt the accuracy of the model. In terms of random forest regressor, this means that ‘pruning’ the tree will generally result in overfitting for my model. Cross-validation method was in line with what was expected and further proved the validity of the dataset. Cross validating between the test dataset and training dataset increased lower the MSE from 0.2045 to 0.1965.



```
def train_and_evaluate(X_train, X_test, y_train, y_test):
    # Train Random Forest Regressor
    rf = RandomForestRegressor(random_state=42)
    rf.fit(X_train, y_train)

    # Predict on test set
    y_pred = rf.predict(X_test)

    # Evaluate performance
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    return mse, r2

# Evaluate for alpha = 0
mse_0, r2_0 = train_and_evaluate(X_train_0, X_test_0, y_train_0, y_test_0)

# Evaluate for alpha = 10
mse_10, r2_10 = train_and_evaluate(X_train_10, X_test_10, y_train_10, y_test_10)

print(f"Alpha = 0: MSE={mse_0:.4f}, R²={r2_0:.4f}")
print(f"Alpha = 10: MSE={mse_10:.4f}, R²={r2_10:.4f}")
```

```
Alpha = 0: MSE=0.2045, R²=-0.0257
Alpha = 10: MSE=0.2070, R²=-0.0383
```

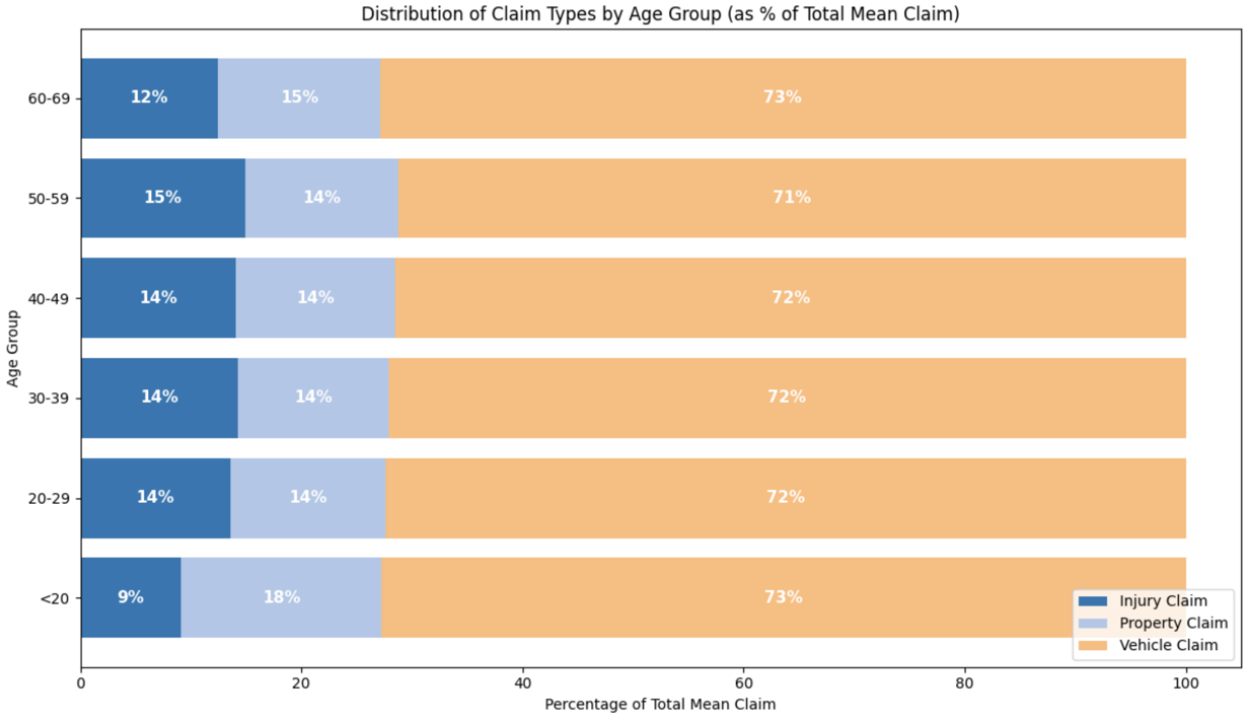
```
def perform_cv(X, y, cv=5):
    rf = RandomForestRegressor(random_state=42)
    mse_scores = -cross_val_score(rf, X, y, cv=cv, scoring='neg_mean_squared_error')
    r2_scores = cross_val_score(rf, X, y, cv=cv, scoring='r2')
    return mse_scores.mean(), r2_scores.mean()

# CV for both datasets
cv_mse_0, cv_r2_0 = perform_cv(X_0, y)
cv_mse_10, cv_r2_10 = perform_cv(X_10, y)

print(f"CV – Alpha = 0: MSE={cv_mse_0:.4f}, R²={cv_r2_0:.4f}")
print(f"CV – Alpha = 10: MSE={cv_mse_10:.4f}, R²={cv_r2_10:.4f}")
```

```
CV – Alpha = 0: MSE=0.1965, R²=-0.0628
CV – Alpha = 10: MSE=0.1977, R²=-0.0684
```

I then explored the grid search cross validation method to see which general set of parameters performed best for my model. Overall, larger and more complicated parameters continued to improve the error of the model, with the best parameters from the samples being at 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200. Continuing the project.



The main question that the provider of the data wanted to answer was the same as mine. “Can we determine whether or not a claim is fraudulent based on the other information surrounding the claim?” So far our dataset proved to have enough information to answer this question. While classification features like fraud reported performs less than optimally with linear regression and mean squared error. The results discussed above showed promising results that we can use to generate a guiding tool for the insurance companies to signal them on how likely the claim is fraudulent. But in further exploring my dataset I will be using tool more fit for classification problems like logistic regression and k-fold cross validation. *Gong (* 2022)

Team Contribution

My team members offered thoughtful and meaningful opinions on my homework. Which helped me better grasp the overall deficiencies in my work. We did not have a communication requirement but met on Discord to help each other out and they submitted their homework in time which helped me cross validate my answer with theirs and gain insights on which methods they used to complete each of the homework. They commented on my work in time to provide feedback before starting the next set of work which gave me ample time to explore the opportunities suggested by my teammates.

Grading:

Moussa Hatab	5/5
Laura Mills	5/5
Rashik Hossain	5/5

Module C Capstone Proposal

Analysis completed on the current dataset during this module provided me with results that enabled me to confide in the current dataset for the questions I wanted to answer for my final project. Along with the main question I wanted to answer “Can we predict how likely a claim is going to be fraudulent,” it also offered insights into other interesting relationships like older customers being likely to be involved in higher paying claims. Which is an important question to be answered for insurance companies as claim payout plays an important role on how they price the policies. In my goal of building a full-feature fraud detection software, this auxiliary information will also provide value for my model to the insurance companies. As previously discussed, the current techniques used for the homeworks were not best suited for binary classification models. For my final project and ultimate goal of building a full-feature fraud detection model - I will be employing techniques more suited for binary classification problems like logistic regression, support vector machine algorithm, and k-fold cross validation instead of the regular cross validation techniques. Formerly, I will be building a fraud prediction model for the Automotive Insurance Industry for their investigation units. Ultimately identifying the fraud risk rating of a claim using the information surrounding the claim.

References

Aqqad, A. (2023, August 22). *Insurance_claims*. Mendeley Data. https://data.mendeley.com/datasets/992mh7dk9y/2

FBI. (2010, March 17). *Insurance fraud*. FBI. https://www.fbi.gov/stats-services/publications/insurance-fraud

Björklund, M. (2019). Be careful with your principal components. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.03440/

Centers for Disease Control and Prevention. (n.d.). *Older adult drivers*. Centers for Disease Control and Prevention. https://www.cdc.gov/older-adult-drivers/about/index.html#:~:text=Age%2C%20sex%2C%20and%20age%2D,to%20injury%20in%20a%20crash.

Nau, R. (2022). What's a good value for R-squared? https://people.duke.edu/~rnau/rsquared.htm