Insurance Market Segmentation in Panama City by K-Prototype Clustering

DREXEL LEBOW CAPSTONE PROJECT

FALL 2024

Group 4
Aabhasi Chachire, Ananya Shankar, Anirudh Shivaram, Mathieu Ndong, Thu Thai

EXECUTIVE SUMMARY

Florida’s residential insurance market, particularly in Panama City, is under immense strain due to the rising frequency and intensity of natural disasters such as hurricanes and flooding. Despite homeowners paying some of the highest premiums in the nation, insurers face unsustainable claim costs, forcing many to withdraw from the market. The result is a destabilized industry with systemic inefficiencies and limited options for risk management. Addressing these challenges requires innovative, data-driven solutions to restore profitability and stability while adapting to the region’s high-risk environment.

To tackle this issue, our team utilized a comprehensive dataset encompassing flood and coastal risks, demographic characteristics, and property attributes to develop a robust clustering model. Leveraging **K-Prototypes Clustering**, which integrates both numerical and categorical data, the team identified key factors influencing premiums, including flood risk score, property age and size, demographics, and distance to the coast. The model segmented the market into four distinct clusters, each representing unique risk profiles and demographic characteristics.

Insights from the clustering analysis provide a strategic framework for insurers to refine their portfolios. For high-risk clusters such as **Cluster 1** (small, modern properties in high-flood areas with a mixed population) and **Cluster 2** (large, modern properties in high-flood areas with an older population), insurers should apply higher premiums and deductibles to offset claim potential and rebuilding costs. For **Cluster 3** (smaller, older properties with moderately high flood risk) slightly reduced premiums are appropriate, reflecting the lower likelihood of claims. Besides, **Cluster 4**, characterized by moderate flood risk, newer properties, and a younger population, offers insurers significant growth potential with lower exposure to catastrophic claims. By strategically reallocating resources and developing tailored products that reflect the nuanced risks of each cluster, insurers can optimize their underwriting processes, and profitability.

This data-driven approach empowers insurers to mitigate exposure to high-risk properties while fostering balanced growth. It also highlights opportunities for competitive differentiation through targeted pricing strategies and innovative offerings, such as discounts for preventative measures. The proposed solution not only enhances operational efficiency but also positions insurers for long-term success in Florida’s challenging and evolving insurance landscape.

PROBLEM STATEMENT

The insurance industry Panama City faces unprecedented challenges due to the increasing frequency and severity of natural disasters, including hurricanes and coastal flooding during storm seasons. These risks have pushed many insurers to a breaking point, as claims from property damage continue to outpace revenue generated from premiums. Despite homeowners in Florida paying some of the highest premiums in the nation, with 20% of residents incurring annual costs of $4,000 or more according to realtor.com, insurance companies are struggling to maintain profitability (“Florida Homeowners Are Paying Some of the Highest Insurance Premiums in the Nation”). The high premiums, rather than being a solution, are a symptom of deeper systemic issues in managing risk. Insurers find themselves unable to cover escalating claim costs effectively, leading many to drop policies, such as the case of AAA Insurance, or many to withdraw from the state entirely ("AAA Drops Thousands of Florida Insurance Policies, Citing Risks"). This exodus of insurance providers has left the market in disarray, increasing the financial burden on remaining companies and further destabilizing the industry. The unsustainable nature of current pricing models and the limited options for spreading risk demand innovative approaches to help insurers adapt to these evolving challenges.

DATA REVIEW

Data Selection

To ensure accurate analysis, we focused on residential properties, which constitute the majority of our portfolio. Using the Precisely API queries, we extracted data based on flood risk, coastal risk, demographic characteristics, and property attributes. This comprehensive dataset formed the foundation for our modeling efforts, enabling targeted insights into property-level risks.

Data Cleaning

Our data cleaning process involved importing and consolidating datasets from multiple CSV files, including address fabrics, demographic, property attributes, coastal risk and flood risk data. To ensure data quality, we employed techniques such as duplicate detection and removal, retaining only unique records for consistency. We handled missing values, standardized data formats, and verified the integrity of the data across datasets to align with analytical requirements. Outliers were addressed where necessary to minimize their impact on the model. These preprocessing steps, combining de-duplication, standardization, and validation, ensured that the datasets were clean, consistent, and reliable, forming a strong foundation for accurate and robust model training.

DATA ANALYSIS

Identify key features to Premiums

For our analysis, we started with linear regression. The results revealed minimal to no linear relationships between the dependent variable and most independent variables. Linear regression is best suited for scenarios where there's a clear linear relationship between the independent and dependent variables; hence we explored alternative modeling techniques. We utilized 13 variables derived from a mock insurance quote from Clovered insurance agency website ("Homeowners Insurance in Panama City Beach, Florida"). To identify the most influential factors, we employed random forest and gradient boosting models. Random forest emerged as the superior model, providing most accurate and insightful results. The top 6 variables, identified as the most influential factors by the random forest model, were selected for clustering purposes. *(Appendix 1)*

Initial variables

Target Variable - Premiums

Property Attributes Data - Property living square footage, Roof cover type, Property year built, Building type description, Exterior walls description, floor type

Demographics Data - Property tenure variable, Age group, Income group

Flood risk Data - Flood zone, Elevation, 100-year flood zone distance

Coastal Risk Data - Distance to the nearest coast

Variables used for clustering

Property living sq ft, Elevation, Distance to nearest coast, Property year build, 100-year flood zone distance, Age group

Segment the market by the selected key features

Flood Risk Score Development

To better understand flood risks in Panama City beyond the FEMA flood zones in Bay County, we developed a Flood Risk Score combining two critical factors that we gained from Random Forest model: Address Elevation and Distance to Flood Zone.

The formula for Flood Risk Score is as follows:

$$\textbf{Flood Risk Score} = \textbf{w1} \times \left(1 - \frac{\textbf{Elevation}}{\textbf{Max Elevation}}\right) + \textbf{w2} \times \left(1 - \frac{\textbf{Distance}}{\textbf{Max Distance}}\right)$$

where:

$$w1: Weight\ for\ elevation = \frac{Importance\ Score\ of\ Elevation}{Importance\ Score\ of\ Elevation + Importance\ Score\ of\ Flood\ zone\ distance} = \mathbf{0.55}$$

$$w2: Weight\ for\ distance = \frac{Importance\ Score\ of\ Elevation}{Importance\ Score\ of\ Elevation + Importance\ Score\ of\ Flood\ zone\ distance} = \mathbf{0.45}$$

Weights for elevation (0.55) and distance (0.45) are computed based on their relative importance scores from the Random Forest model. These weights align with the observation that in low-lying areas like Panama City, even small elevation changes significantly affect flood risk. This metric provides a balanced and comprehensive assessment of flood risk.

Why do we choose Clustering Model?

To segment the Panama City market by exposure to high-risk areas, we used clustering to group properties into distinct segments based on key features. This segmentation supports actionable recommendations for insurance companies regarding risk management.

K-Prototypes Model

Regarding clustering models, K-Means is designed for numerical data, using Euclidean distance, while K-Modes specializes in categorical data with Hamming distance. Since our dataset includes both numerical variables (Flood Risk Score, Living Square Footage, Year Built, Distance to Nearest Coast) and a categorical variable (Adult Age Group), we opted for the more advanced K-Prototypes model. This approach combines the strengths of K-Means and K-Modes, making it ideal for handling mixed data types and delivering precise clustering outcomes.

Identify optimal number of clusters by Elbow Method

The Elbow Method was applied to determine the optimal number of clusters for segmenting the data. The graph above illustrates the relationship between the number of clusters and the cost (within-cluster sum of squares). A noticeable "elbow" occurs at n=4, where the reduction in cost begins to diminish. This indicates that using four clusters provides a balance between minimizing intra-cluster variance and avoiding over-segmentation, making n=4 the optimal choice for clustering in this analysis. *(Appendix 2)*

Result of Clusters

The table summarizes the key characteristics of the four clusters identified in the analysis. The values shown represent the mean statistics for each cluster, highlighting differences in flood risk levels, property characteristics (size and year built), and population demographics.

Cluster	Flood Risk	Distance to Coast	Property Size	Year Built	Population Demographics
Cluster 1	High (0.85)	Close (5534 ft)	Small (1399 sq ft)	Modern (1995)	28.8% Above 50 37.9% Below 50
Cluster 2	High (0.88)	Close (5474 ft)	Large (2874 sq ft)	Modern (1993)	51.1% Above 50 23.4% Below 50
Cluster 3	Moderately High (0.79)	Close (5046 ft)	Small (1260 sq ft)	Older (1954)	30.9% Above 50 29.2% Below 50
Cluster 4	Moderate (0.66)	Far (33,428 ft)	Medium (1791 sq ft)	Newer (2005)	19.9% Above 50 49.4% Below 50

WHY CLUSTERING IS BETTER THAN ZIP CODE APPROACH?

Our analysis identified two properties representing distinct risk clusters:

Address 1: 2704 W 20th St, Panama City, FL 32405 (High-Risk Cluster 1)

Address 2: 2816 Krystal Leigh Ct, Panama City, FL 32405 (Moderate-Risk Cluster 4)

The first property is vulnerable due to its proximity to the coast and low elevation, while the second is in a low-risk cluster. Despite similar characteristics—three bedrooms, two bathrooms, and identical materials—the low-risk property has a premium twice as high as the high-risk one.

Surprisingly, these homes are in the same county, zip code, and only five miles apart, highlighting flaws in zip code-based premium calculations.

This demonstrates the value of our clustering model, which factors in elevation, risk proximity, and other nuances, enabling insurers to set premiums that are both accurate and fair.

BUSINESS RECOMMENDATIONS

For Cluster 2, comprising high flood risk, large, modern properties, and an older population, the highest premiums are necessary due to the higher claim potential. Cluster 3, with moderately high flood risk, smaller, older properties, and a mixed population, should have slightly lower premiums reflecting reduced claim potential. Cluster 1, with high flood risk, small, modern properties, and a mixed population, also warrants high premiums to account for rebuilding costs and flood risk. Cluster 4, with moderate flood risk, medium-sized, newer properties, and a younger population, should have competitive, moderate premiums due to its lower risk profile and newer construction.

With around 60% of the portfolio in high-risk clusters (combination of 1 and 2), managing risk exposure is critical. The client should prioritize expanding their exposure to Cluster 4 as this is the lowest risk cluster, also with the smallest proportion in their portfolio. This cluster offers lower risk, newer properties, and a younger, proactive population. By offering competitive premiums and innovative products like discounts for flood-proofing or home automation, insurers can attract this promising market while minimizing risk. Besides, with the current status of overexposure to high-risk areas, the client should thoroughly reassess policies in Cluster 1 and Cluster 2, considering dropping policies, or increasing deductibles for policies with a high likelihood of claims.

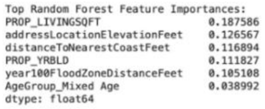
REFERENCES

“AAA Drops Thousands of Florida Insurance Policies, Citing Risks." CBS News, 21 Nov. 2024, https://www.cbsnews.com/news/aaa-insurance-policies-florida-nonrenewal/. Accessed 22 Nov. 2024.

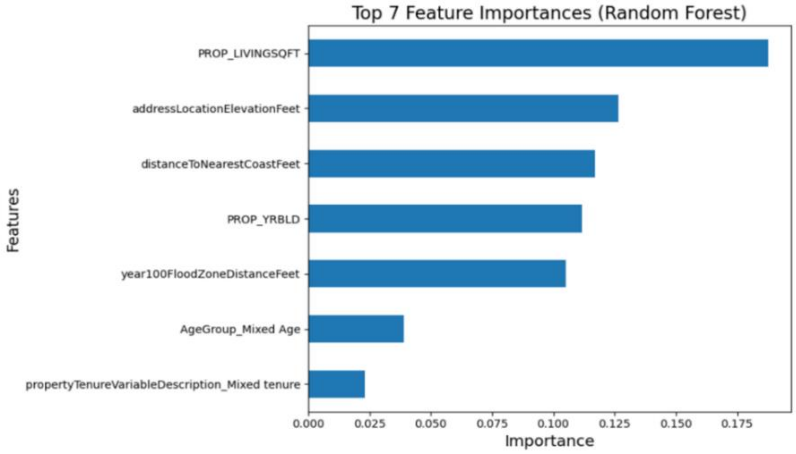
"Florida Homeowners Are Paying Some of the Highest Insurance Premiums in the Nation." *Realtor.com*, Realtor, www.realtor.com/news/trends/florida-home-insurance-cost-premiums/. Accessed 20 Nov. 2024.

"Homeowners Insurance in Panama City Beach, Florida." Clovered, https://clovered.com/homeowners-insurance/florida/panama-city-beach/. Accessed 22 Nov. 2024.

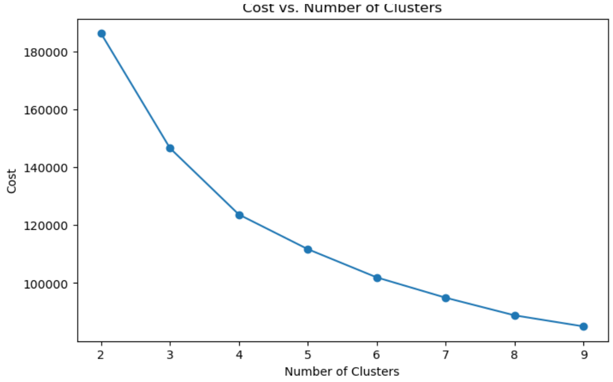
APPENDICES



```
Top Random Forest Feature Importances:
PROP_LIVINGSQFT                   0.187586
addressLocationElevationFeet      0.126567
distanceToNearestCoastFeet        0.116894
PROP_YRBLD                        0.111827
year100FloodZoneDistanceFeet      0.105108
AgeGroup_Mixed Age                0.038992
dtype: float64
```



Appendix 1



Appendix 2