**AIT 622 - Big Data Needs Analysis**

**AIT622-Team Project Instructor: Dr. Hadi Rezazad**

**Part 2:**

**“Project Summary and Findings” Team-6**

**Narasimha Krishna Amruth Vemuganti - G01389960**

**Kaushik Prayaga - G01354233**

**Tharuni Myda - G01393733**

**Jaswanth Erusu- G0137621611**

**Apr 26, 2023**

**a) Assume you are responding to the collecting organization’s RFP (Request for Proposal)**

Response to FEMA's National Flood Insurance Policy Database's Request for Proposals

The dataset holds tremendous promise for a variety of applications, in our team's analysis of it. The collection includes more than 50 million transactions for flood insurance policies, each with several fields concerning anonymous flood policy holders in the United States. It also contains information on the building's elevation, flood zone rating, zip code, county code, census tract, state code, city code, building type, business type, flood insurance deductibles, total insurance coverage, and total insurance premium of the policy, among other details.

We are aware that the National Flood Insurance Program (NFIP) backs most of the market for offering flood insurance policies in the United States. Only policies purchased inside participating towns are eligible for the NFIP, which collaborates with private insurance providers to offer flood insurance coverage to individuals and businesses. There is a critical need to more effectively distribute financial risk associated with climate change because flooding is the main cause of economic harm to local communities, as shown by the hurricane seasons from 2016 to 2019, and because destructive flooding is predicted to increase because of climate change. To create a thorough method for assessing flood risk, our team suggests utilizing the FEMA National Flood Insurance Policy Database. With the use of this tool, our clients will be able to evaluate possible flood hazards and make decisions that are well-informed. We shall follow the following phases in our strategy:

**Exploratory Data Analysis:** To understand the structure and relationships within the data, we will do exploratory data analysis (EDA). Descriptive statistics, graphics, and grouping analysis will all be a part of our EDA.

**Machine Learning Modelling:** We will create and train machine learning models to forecast flood hazards based on the insights obtained from the EDA. Our models will employ a variety of methods, such as gradient boosting, decision trees, and random forests.

**Risk Assessment Tools :** To create a thorough flood risk assessment tool, we will finally incorporate the machine learning models. Our clients will be able to use the tool to get practical knowledge about potential flood hazards, enabling them to make wise decisions regarding their insurance plans, real estate investments, and risk management techniques.

**b) Identify all the organization’s stakeholders connected to the project.**

**Stakeholders**:

**Federal Emergency Management Agency (FEMA):**

A significant player in this initiative is the Federal Emergency Management Agency (FEMA), a federal organization that oversees the National Flood Insurance Program (NFIP).

**Real estate companies:**

As they may utilize the data to inform their assessments of flood risks when purchasing or selling properties, real estate businesses could be project stakeholders.

**Insurance companies:**

Insurance companies involved with the NFIP use the database to manage their policies, and issue insurances who are affected could be stakeholders in this project.

**Real estate companies:**

Real estate firms could be involved in this effort since they could use the data to help them assess the danger of flooding when purchasing or selling properties.

**Academic institutions:**

Due to the possibility that they would use the data for research and education, academic institutions may be participants in this study.

**c) Are there any privacy, quality, or other issues with this data.**

**Privacy concerns:** The policyholders' privacy and security may be at danger because some of the columns contain personally identifiable information (PII) about them, such as their reported zip code, latitude, and longitude.

**Quality issues:** Problems with the dataset's quality include a sizable number of missing values and inconsistent data across many variables. For instance, there are several missing values in the "basement enclosure crawlspace type" column, which may affect the precision and dependability of any studies or models built using the data.

**ii) Evaluate & recommend build-vs-buy solutions for the project**

For the project including the dataset National Flood Insurance Policy Database of FEMA, it is further important to assess and provide construct vs. purchase options. It is vital to think about using big data technologies like Apache Hadoop or Apache Spark for distributed computing given the dataset's size, which includes more than 50 million national flood insurance policy transactions. Additionally, Python libraries like NumPy and Pandas may be used for pre-processing and cleaning data, while Matplotlib and Seaborn can be used for data visualization.

**d) Requirements: What resources do you need to complete the project?**

**i) People, technology (HW, SW)**

Several resources are needed to finish the endeavour of evaluating the National Flood Insurance Policy Database of the FEMA.

The most important resource needed is people. The size of the project will determine if a team of data scientists, data engineers, and subject matter experts is required. While the data engineers would be in charge of organizing, integrating, and converting the data, the data scientists would be in charge of evaluating the data and drawing conclusions from it. To give context and aid with feature engineering, subject matter experts would be needed. Technology is the second resource that is needed. Both software and hardware are included in this.

**ii) Evaluate & recommend build-vs-buy solutions for the project**

For the project including the dataset National Flood Insurance Policy Database of FEMA, it is further important to assess and provide construct vs. purchase options. It is vital to think about using big data technologies like Apache Hadoop or Apache Spark for distributed computing given the dataset's size, which includes more than 50 million national flood insurance policy transactions. Additionally, Python libraries like NumPy and Pandas may be used for pre-processing and cleaning data, while Matplotlib and Seaborn can be used for data visualization.

**e) Describe the relevant metadata (types of data in the dataset)**

**Metadata**

Distinctive features or characteristics of flood insurance policies and policyholders in the United States are represented by the variables in this dataset. Each field is briefly described below:

**Flood Zone rating-** The flood zone was calculated via the Flood Insurance Rate Map (FIRM) employed for rating the insured building or complex.

**Building Elevation-** Building elevation is a measurement of how high a structure is above either ground level or sea level. Because homes in low-lying locations are more vulnerable to floods compared to those in higher altitudes.

**Zip code, county code, census tract, state code, city code -** geographic markers that specify the assets’ and the policyholder's whereabouts. These codes make it easier to comprehend how local shifts in flood risk affect policyholders and the geographic distribution of flood insurance contracts.

**How the policy was originally rated:** The procedure used to calculate flood coverage cost.

**Census tract-** statistical divisions that are revised before every decennial survey inside a county or comparable group. Geocoding system is used by the NFIP (National Flood Insurance Program) to allocate census tract codes.

knowledge about the variables influencing the price of flood insurance for various kinds of buildings and businesses. This knowledge can then be used to inform policyholders' choices regarding how to reduce their flood risk and safeguard their finances in the case of a flood.

**f) Statistical Analysis**

The appropriate types of statistical analysis that can be performed on the FEMA (Federal Emergency Management Agency) dataset containing both numerical and categorical fields.

**Descriptive Statistics**

The descriptive statistics for the numerical fields include count, mean, standard deviation, minimum, maximum, and quartiles. The descriptive statistics for the category columns, however, provide a count of frequencies and percentages. In the categorical attributes, the frequency measure and proportion for each class are calculated. The numerical fields might be represented by histograms, while the categorical columns are by count plots.

**Correlation Analysis**

Correlation analysis may also be performed on numerical columns by generating the correlation matrix and producing a heatmap with the seaborn library. For example, the total building insurance coverage and the total contents insurance coverage showed a significant positive association, showing that as the total building coverage grew, so did the total contents coverage.

**Regression Analysis**

The correlation between a numerical response parameter and one or more other numeric or categorical predictor variables can be modelled using regression analysis. For instance, we may model the link between the overall insurance premium and the construction insurance coverage and contents insurance coverage using logistic regression, Poisson regression, or linear regression.

Chart

Description automatically generated with low confidence Chart, scatter chart

Description automatically generated

Figure 1. Heat Map for the Correlation Analysis Figure2. Regression analysis plot

**Hypothesis Testing**

Using sample data, this study is conducted to evaluate a population-related hypothesis or assertion. For example, Testing if there is a significant variation in the mean total insurance premium of the policy across homes situated in various flood zones is a hypothesis test that may be performed on the data.

**Define Terms**

**FIRM -** Flood Insurance Rate Map

**FEMA-** Federal Emergency Management Agency

**Numerical Fields**- The features of a dataset that have numerical values.

**Categorical Fields-** The fields of a dataset that contains the data which can be segregated into specific categories.

**Frequency Measure-** A count of how frequently each data appears in a categorical variable.

**Correlation analysis-** A statistical technique that assesses the strength of the connection between two or more variables.

**Heatmap**- Using colours to denote the strength of the correlation between variables.

**Mean-** The average value of a group of numerical observations.

**Variation-** How far a given observation deviates from the mean value.

**h) Estimate the time/effort required for the study**

**Data Cleaning and Pre-processing**: Cleaning and pre-processing the data is the initial stage in every data analysis effort. Depending on the quality of the data, this process might be time-consuming and labour-intensive.

**Data Exploration and Analysis:** After cleaning and pre-processing the data, the next stage is to explore and analyse it. Depending on the study aims, this stage may include techniques such as descriptive statistics, data visualization, machine learning algorithms, and others.

**Statistical Modelling and Inference:** Statistical modelling and inference may be necessary to test hypotheses, make predictions, or draw conclusions, depending on the study objectives. This stage may need complicated statistical approaches and a large amount of time and effort.

**i) anticipated advantages and value for the company in performing the research**

The organization should expect several values/benefits from the research it conducts utilizing the FEMA flood insurance dataset.

**Risk evaluation**

The research can assist the organization in evaluating the risk involved in insuring properties in various flood plains, as well as with various building kinds, commercial enterprise types, and geographic locations. With this information, the company will be better able to assess the risks connected to various kinds of properties and enterprises and take the necessary precautions to reduce such risks.

**Pricing tactics**

The analysis may also bolster the company in improving its flood insurance policy pricing strategies. The company may determine the variables that impact coverage costs by studying the data and then modifying the cost models appropriately.

**Compliance**

The research can assist the firm in ensuring regulatory compliance. The company can determine instances in which they might be out of compliance by reviewing the data and then taking remedial action to remedy any concerns.

**Better ability to make decisions.**

Decision-making will be improved because of the research, which can give the organization a substantial perspective into the elements that contribute to the performance of their flood insurance plan. The organization may make better-informed judgments that will help the program succeed by applying this understanding to guide its decision-making processes.

**Definitions :**

**Request for Proposal (RFP) -** A written description of the specifications, requirements, and scope of work for a project or service that is being sought from possible vendors or service providers.

**National Flood Insurance Program (NFIP) -** NFIP is a government-sponsored insurance program that offers flood insurance to Americans who own homes, rent apartments, and operate businesses.

**Gradient Boosting -** A machine learning method that entails integrating many ineffective models to produce a robust model that can produce reliable predictions.

Decision Trees - A machine learning technique that classifies data using a tree-like model of choices and potential outcomes.

**SQL** - Relational databases are managed and worked on using the computer language SQL.

Tableau - Users may generate interactive and dynamic displays of data sets using the business intelligence and data visualization application Tableau.

**Big Data** - Large, complicated data collections that are challenging to manage, handle, and analyse using conventional data processing techniques are referred to as "big data."

**Apache Hadoop** - Large data sets may be stored and processed in a distributed fashion thanks to the open-source software framework known as Apache Hadoop.

**Apache Spark** - Apache Spark is a distributed computing system that is open-source and made to analyse massive amounts of data fast.

Definitions:

The National Flood Insurance Program (NFIP): NFIP is a government program created by Congress in 1968, offers insurance protection against flood damage to property owners in participating communities. The Federal Emergency Management Agency (FEMA) is in charge of running the program.

Stakeholders: People or organizations who are involved in or affected by the project. They might have a say in project decisions, be impacted by the results, or participate in its execution.

Assessments: Evaluations or studies of the risks of flooding or other pertinent aspects that may have an impact on real estate or insurance industry decision-making.

Personally identifiable information (PII): Information that may be used to identify a specific person, such as their name, address, social security number, or other special identifiers, is known as personally identifiable information (PII).

Privacy Act: Federal law governing the gathering, use, and disclosure of personal data by federal agencies is known as the Privacy Act of 1974. It gives people the right to access and update their personal information that the government has on file.

Geographical coverage: The degree to which the data covers various geographic locations or regions is referred to as geographic coverage. The information in this instance excludes policies from other providers and solely includes policies issued by the National Flood Insurance Program (NFIP).

References:

1. Author. (2019, July 4). *Home*. Wise. Retrieved April 24, 2023, from <https://wise-answer.com/who-are-the-stakeholders-in-fema/>
2. *Guidance for stakeholder engagement - FEMA*. (n.d.). Retrieved April 24, 2023, from https://www.fema.gov/sites/default/files/2020-02/Stakeholder\_Engagement\_Introduction\_Guidance\_Nov\_2019.pdf
3. Kaggle dataset description: https://www.kaggle.com/datasets/lynma01/femas-national-flood-insurance-policy-database
4. FEMA privacy policy: <https://www.fema.gov/privacy-policy>
5. *FEMA’s National Flood Insurance Policy Database*. (2020b, May 29). Kaggle. <https://www.kaggle.com/datasets/lynma01/femas-national-flood-insurance-policy-database>
6. *OpenFEMA*. (n.d.-b). FEMA.gov. <https://www.fema.gov/media-library/assets/documents/180376>
7. Frank, T. (2022, September 15). *FEMA removes data showing drop in flood insurance policies*. E&E News. <https://www.eenews.net/articles/fema-removes-data-showing-drop-in-flood-insurance-policies/>