household electric bill consumption

project done by

Subhajyoti Maity

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah
Pin - 711 202, West Bengal

Contents

1 Introduction:

The dataset includes information on several variables, including the number of people residing in each household, the size or square footage of the homes, and the corresponding monthly electricity consumption.

The analysis of the dataset, which focuses on the relationship between the number of people,use of T.V and A.C,income and home size with monthly electricity consumption in the household, aims to gain insights into how these factors influence energy usage.

It allows homeowners to identify potential areas for improvement and implement energy-saving measures tailored to their specific household characteristics.

To conduct the analysis, statistical techniques such as regression analysis, correlation analysis, or data visualization methods can be employed. These approaches can help uncover insights, quantify relationships, and visualize the impact of the number of people and home size on monthly electricity consumption.

Machine learning algorithms are commonly employed to build predictive models for household electric bill consumption. These models are trained on historical household data, where the area of house and corresponding features are known. The algorithms learn patterns and relationships within the data to make accurate predictions on unseen or future household instances.

2 Data collection

I collected the data from the site
`https://www.kaggle.com/datasets/gireeshs/household-monthly-electricity-bill`

3 Data Description:

Dataset has 1000 rows and 10 columns. All values are non-null. Data type of most of the columns is int64. whereas housearea,`ave_monthly_income` and `amount_paid` are in float64.No of Discrete Values Numeric Columns is 7 and others three are continuous value numeric column. some of the columns' description given below:

- `num_rooms`:Number of room in the particular house.
- `num_people`:Number of people in the house
- housearea:Area of the house in square foot.
- `ave_monthly_income`:Average monthly income of the household.
- `num_children`:Number of children in the particular house.
- `is_ac`:Is AC present in the house?
- `is_tv`:Is TV present in the house?
- `is_flat`:Is house a flat?
- `is_urban`:Is the house present in an urban area?
- `amount_paid`:Amount paid as the monthly bill.

4 Data Preprocessing:

Data preprocessing is an essential step in the data analysis and machine learning pipeline. It involves cleaning, transforming, and organizing raw data into a format suitable for further analysis or model training.
The process of data preprocessing typically includes the following steps:

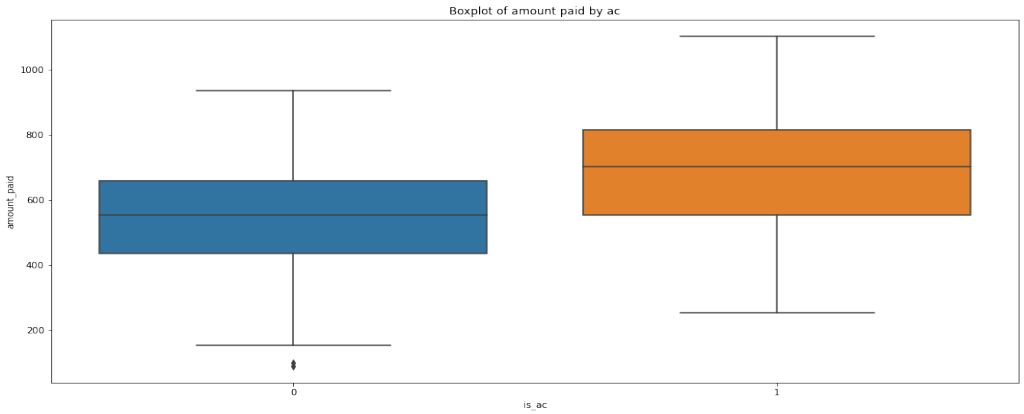
- Data Cleaning:This step involves handling missing values, noisy data, and inconsistencies in the dataset. Missing values can be filled or removed based on various techniques, such as mean imputation, median imputation.But in our dataset there does not exist any missing value.
- Handling Imbalanced Data:Some column in the dataset contain unnecessary value such as number of rooms and number of people of a family can not happened -1.So i remove those rows from the dataset which contain -1 value in columns `num_rooms` and `num_people`.

After removing those rows from the dataset which contain -1 value in columns num_rooms and num_people ,new Dataframe has 991 rows and 10 columns.we do our Exploratory Data Analysis over new dataframe.

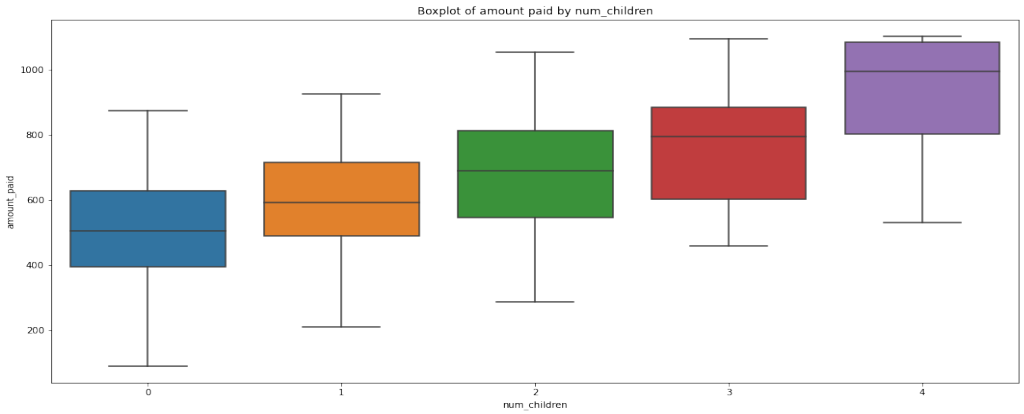
5 Exploratory Data Analysis(EDA):

EDA is a critical phase that involves visualizing and analyzing the dataset to gain insights into the distribution of features and their impact on household electic amount paid .
EDA helps to uncover insights, identify data quality issues, and guide further analysis or modeling tasks. Create visualizations to explore and understand the data better. Commonly used plots include histograms, box plots, scatter plots and bar plots.

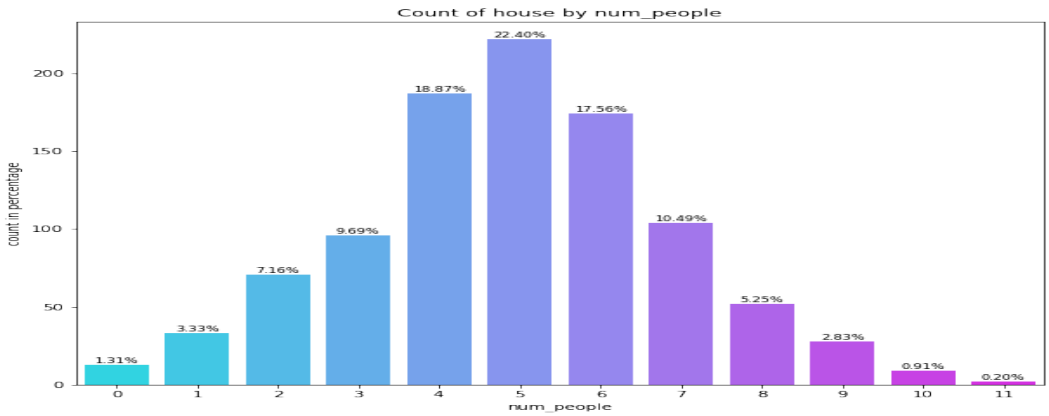
- some following insights are given below:



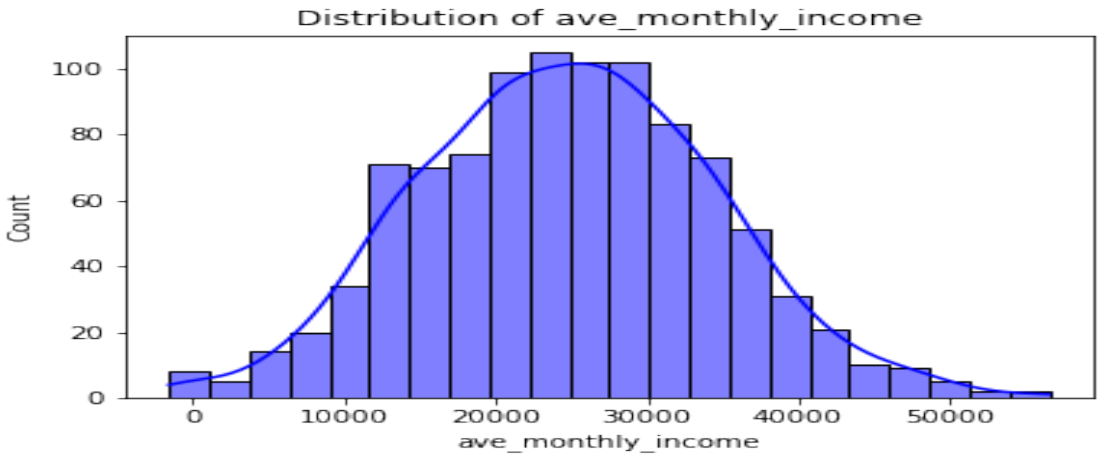
from the above picture we can say that The dots above or below the whiskers represent individual data points that are considered outliers, falling outside the typical range of amount paid of house bill. Outliers may indicate specific house with significantly higher or lower amount paid of house bill compared to the rest of the house bill.By comparing the positions and distributions of the boxes, we can observe amount paid of house bill differences among thethe houses.while use of ac seems to be the most expensive amount paid of house bill.



from the above picture we can say that By comparing the positions and distributions of the boxes, we can observe amount paid of house bill differences among the number of children of each houses.while the house where number of children is 4 be the most expensive amount paid of house bill.



from the above picture we can say that By comparing the positions and distributions of the boxes, we can observe most of the house have number of people is 5 in the given dataset.



from the above picture we can say that most of the family income in the dataset lies between 23000 and 26000.

6 Split the data:

- The scikit learn algorithms take two separate arguments. This means they need independent variables separately and the dependent variable (or target variable) separately. But since in the train dataset both independent and dependent variables are present together so I need to separate them out.
- Firstly, I'll create a set of independent variables from the train dataset. So I'm dropping the 'target' variable from it using axis=1. This axis=1 specifies that the drop shall happen from the column.
- The data set will be split into 80% train and 20% test.

7 model selection:

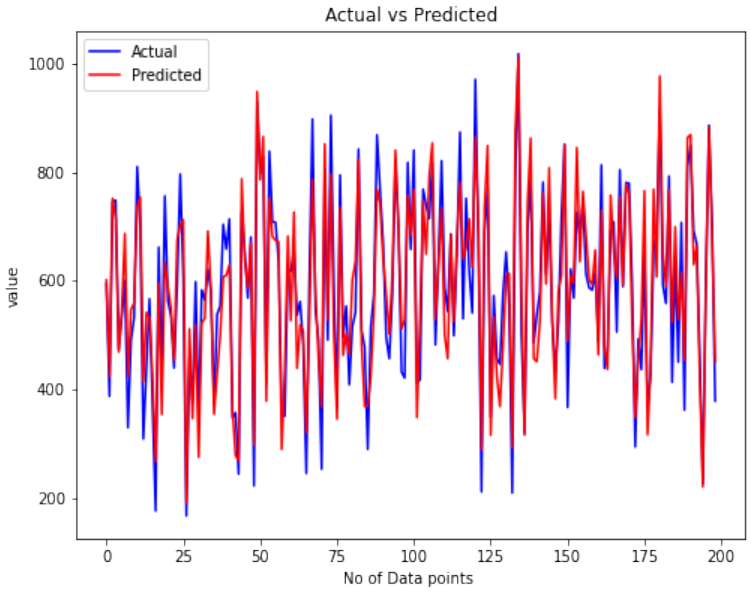
We employed several machine learning models to predict amount paid of electric bill of each house. The models used in this study are as follows:

1. Linear Regression
2. Ridge Regression

8 Performance Evalution

1. Linear Regression : We fit the whole dataset into linear regression model. we get the value of R2 score is 0.8625962249106472.

2. Ridge Regression : We fit the whole dataset into Ridge regression model. we get the value of R2 score is 0.8628401832503297 .



9 Conclusion:

However, it is crucial to recognize the limitations of the predictive model. The accuracy of bill consumption predictions may be influenced by factors not present in the dataset, such as seasonal variations, weather conditions, or lifestyle changes. Therefore, continuous monitoring and updates to the model are recommended to ensure its accuracy over time.

In summary, the household energy bill consumption prediction analysis has shed light on the dynamics between household characteristics and monthly electricity consumption.