Market research survey

vinusha gorijala- CS5300

May 4, 2023

Contents

1 Introduction

The database to which you are referring is a dataset from a market research study that tries to compile demographic data on consumers' preferences for computer brands. Seven tables make up the dataset: brand, salary, age, elevel, car, zipcode, and car. The respondents' annual incomes are shown in the pay table, while their ages are shown in the age table. The respondents' education level is shown in the elevel table, and whether they possess a car is shown in the car table. The respondents' zip codes are listed in the zipcode table, while their credit scores are listed in the credit table. The brand table also includes data on the respondents' preferred computer brands. This dataset may be utilized for a wide range of market research and data analysis tasks, including spotting trends in consumer brand preferences for computers and figuring out how demographic factors affect these choices. Additionally, it may be utilized for machine learning applications, such as creating predictive models that aid businesses in creating more successful marketing plans.

2 overview

The market research survey dataset includes two files: CompleteResponses.csv and SurveyIncomplete.csv. The former is the training split that contains nearly 10,000 answered surveys, while the latter is the test split used for further analysis. The dataset comprises seven tables: salary, age, elevel, car, zipcode, credit, and brand. The salary table shows the respondents' annual income, the age table displays their age, and the elevel table indicates their education level. The car table indicates whether respondents own a car or not, and the zipcode table provides information on their zip codes. The credit table shows the respondents' credit score, while the brand table provides data on their computer brand preferences. The dataset can be used for market research purposes, data analysis, and machine learning applications. It can help companies gain insights into consumer preferences for computer brands and develop effective marketing strategies. The dataset also allows for understanding how demographic factors influence computer brand preferences and can be used to build predictive models that inform marketing campaigns. Overall, this dataset is a valuable resource for businesses seeking to understand their customers and improve their marketing efforts.

3 Dataset

I downloaded the dataset from Kaggle, which is a good repository of datasets. The dataset contains the following columns:

- **Salary**: The represent the salary contains the annual salary of the respondent in USD.
- **Age**: Th represent the age of the respondent in years.
- **Elevel**: The represent the education level of the respondent. The education level is divided into six categories: High School, Some College, Bachelor's Degree, Master's Degree, Doctorate, and Professional Degree.
- **Car**: The represent information on whether the respondent owns a car or not. The values in this column are "Yes" and "No".
- **Zipcode**: The represent the zip code of the respondent's location.
- **Credit**: The represent the credit score of the respondent.
- **Brand**: The represent the computer brand preference of the respondent. The values in this column are "Apple", "Dell", "HP", "Lenovo", "Microsoft", "Samsung", and "Other".

3.1 Visualization of the distribution of each input features

The below histograms shows the each info highlights of how they distributed before normalization.

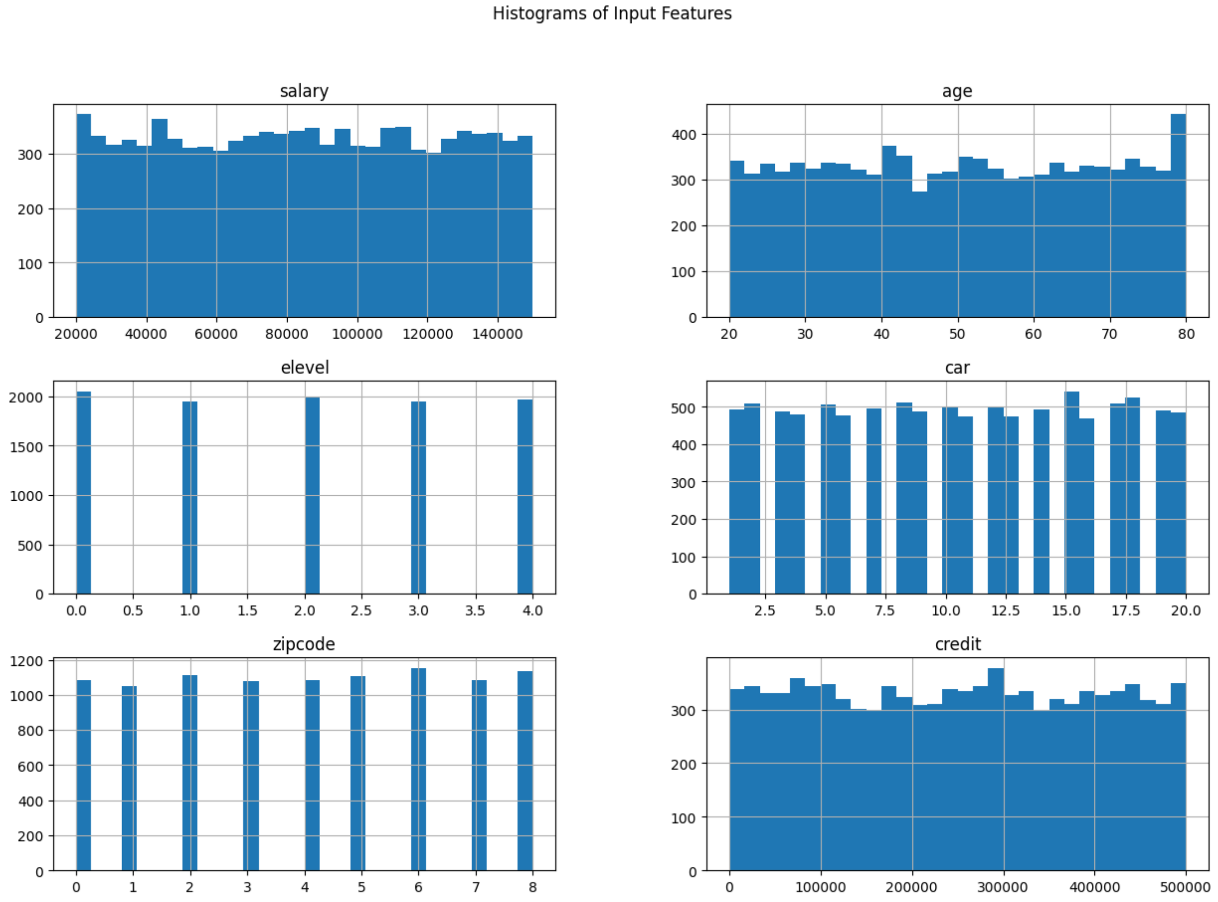


Figure 1: Input Data Distribution Histograms

3.2 Distribution of the output labels

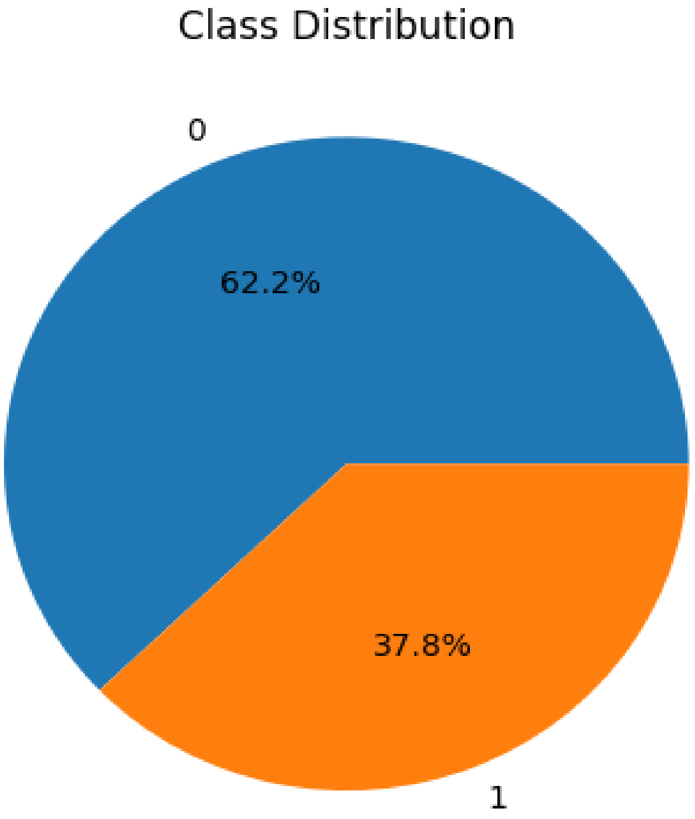


Figure 2: Output Data

4 Data Processing

4.1 Data Splitting

The data was randomly moved, and the set of data was divided into training and validation sections, with 80

4.2 Data Normalization

To handle the variance in the distribution of data, data preparation is required before data mining. To do this, normalizing methods are utilized to make the optimization problem more numerically stable and to enhance training. Normalization ensures that all values are between 0 and 1, and that outliers appear within the normalized data. There are two normalization strategies accessible, each with its own set of effects, although either methodology can be employed for the time being.

Mean Normalization Formula

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Z-Score Normalization

$$X_{normalized} = \frac{X - X_{mean}}{X_{standard_deviation}}$$

To center the table in Overleaf, you can use the following code:

salary	age	elevel	car	Zipcode
credit	brand			
119806.54480	45	0	14	4
442037.71130	0			
106880.47840	63	1	11	6
45007.17883	1			
78020.75094	23	0	15	2
48795.32279	0			
63689.93635	51	3	6	5
40888.87736	1			
50873.61880	20	3	14	4
352951.49770	0			
87580.91422	75	1	18	8
282511.90950	1			
282511.90950	75	2	7	4
384871.36390	1			
97828.08884	66	2	15	0
399446.69620	1			
20000.00000	24	1	14	1
223204.64950	1			
96430.16419	34	1	2	7
224029.80700	0			

Input Feature Statistics before Normalization

Salary	age	elevel	car	Zipcode	credit	brand
0.767743	0.416667	0.00	0.684211	0.500	0.884075	0.0
0.668311	0.716667	0.25	0.526316	0.750	0.090014	1.0
0.446313	0.050000	0.00	0.736842	0.250	0.097591	0.0
0.336076	0.516667	0.75	0.263158	0.625	0.081778	1.0
0.237489	0.000000	0.75	0.684211	0.500	0.705903	0.0
0.519853	0.916667	0.25	0.894737	1.000	0.705903	0.0
0.839857	0.916667	0.50	0.315789	0.500	0.769743	0.1
0.598678	0.766667	0.50	0.736842	0.000	0.798893	1.0
0.000000	0.066667	0.25	0.684211	0.125	0.446409	1.0
0.587924	0.233333	0.25	0.052632	0.875	0.448060	0.0

Input Feature Statistics after Normalization

4.3 Normalized Data

The histogram plot after normalization of each info highlights indicating their most extreme and least value as well as how they are distributed can be found in the pictures given underneath.

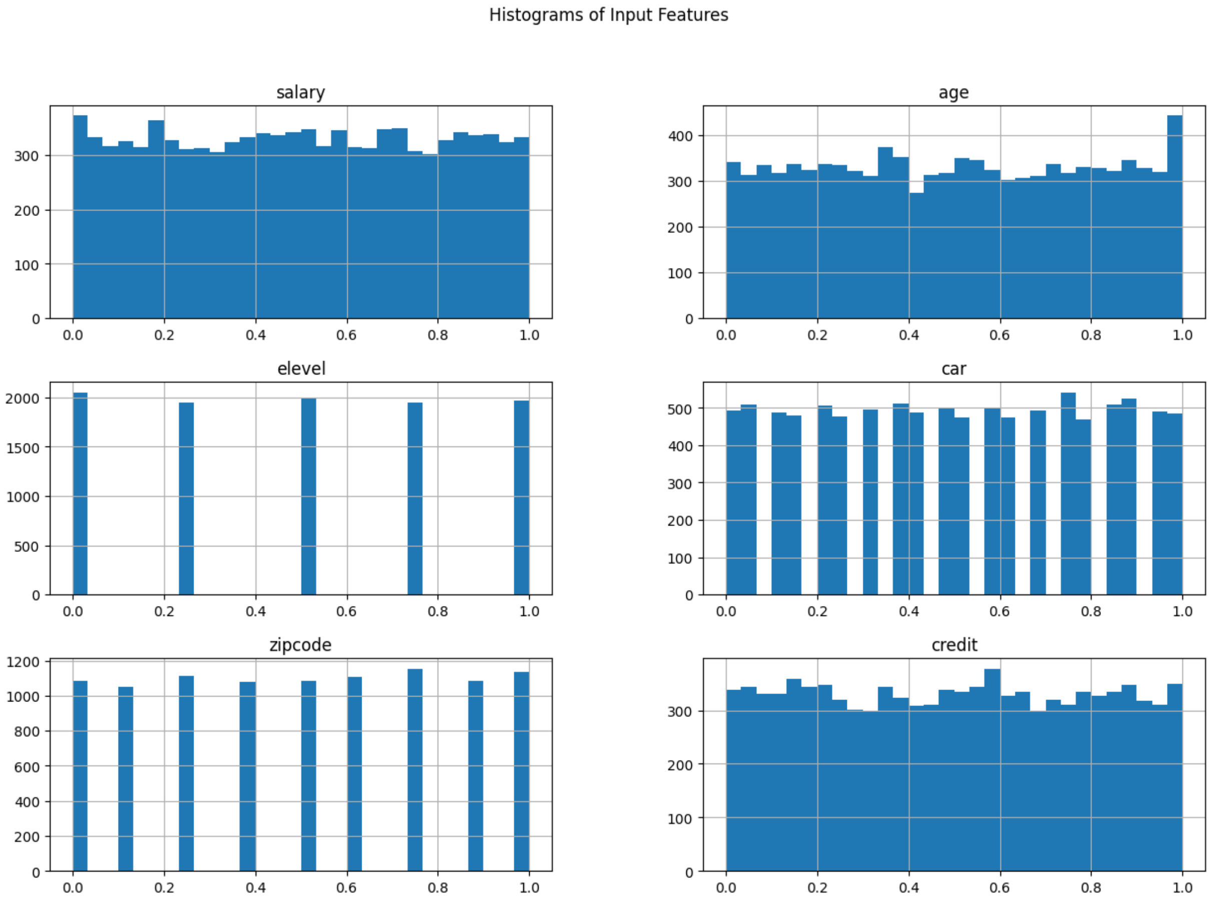


Figure 3: Input Data Distribution Histograms - Before Normalization

5 Modelling

A feed forward artificial neural network architectures was used to create the model.

NOTE: Data is shuffle. Thus, the result will vary every time. All models were compiled and fit on May 2, 2022.

5.1 Selected Neural Network Architecture

5.1.1 Single layer model

The first model is a baseline model (which serves as a control) with a simple architecture consisting of one input and one output layer. It will then steadily rise by one secret layer, eventually reaching a maximum of five hidden levels. During the training, an early preventing strategy was employed. Train accuracy: 0.6167, Validation accuracy: 0.6247

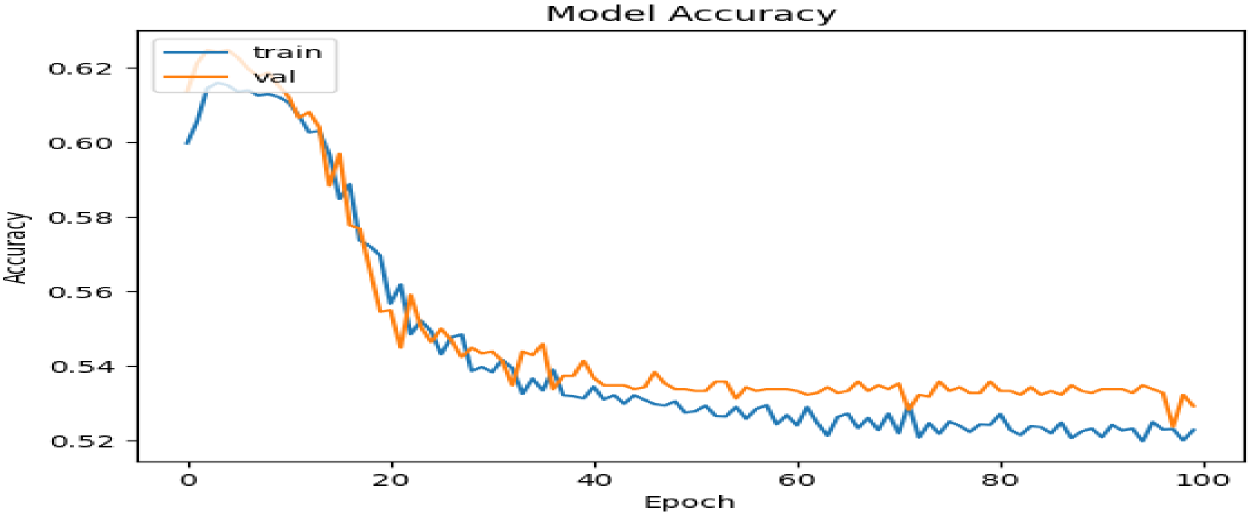


Figure 4: curve showing model loss

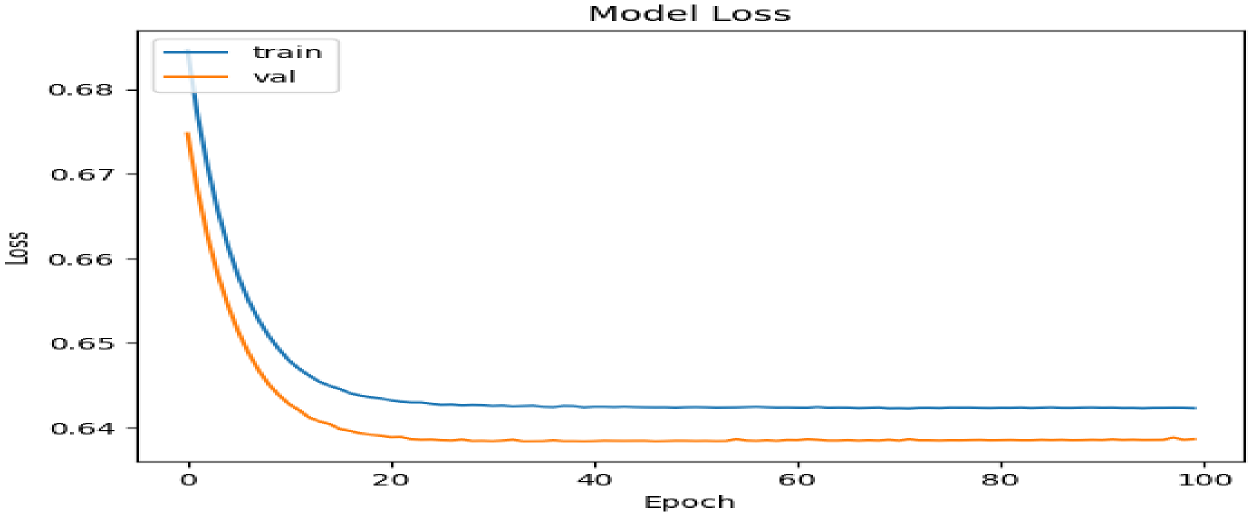


Figure 5: curve showing model loss

5.1.2 Multilayer

On the given training and validation datasets, the deep neural network model with four dense layers and neurons was built using the binary cross-entropy loss and Adam optimizer. The best models for training and validation accuracy are stored during training and then loaded for evaluation. The accuracy and loss of the model are displayed for both the training and validation datasets. Finally, the best model on the validation dataset is used to produce the F1 score, precision, and recall measures. Train accuracy: 0.9614, Validation accuracy: 0.9121

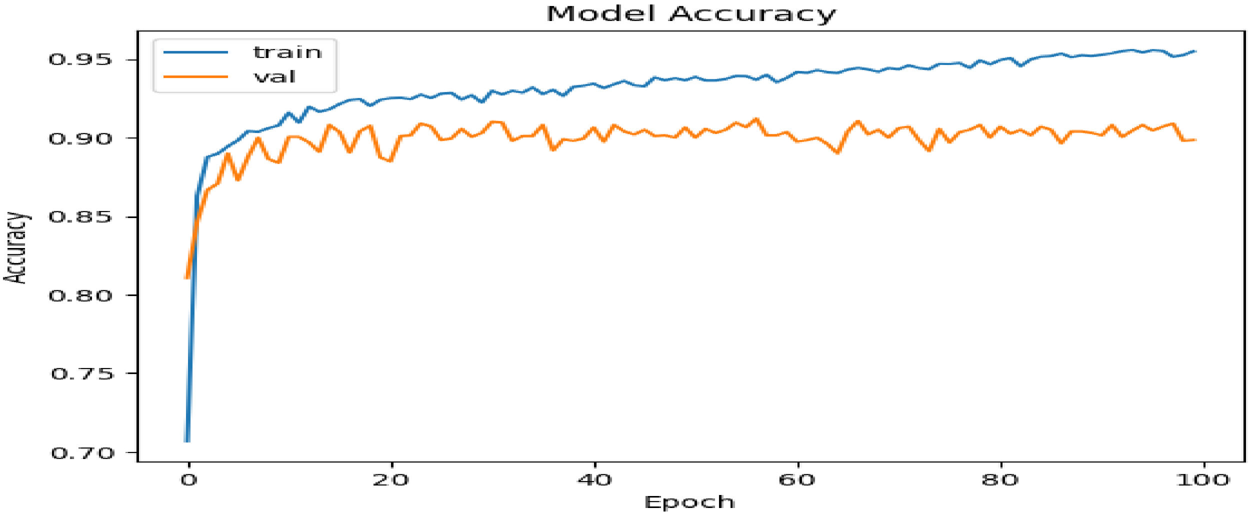


Figure 6: curve showing multilayermodel accuracy

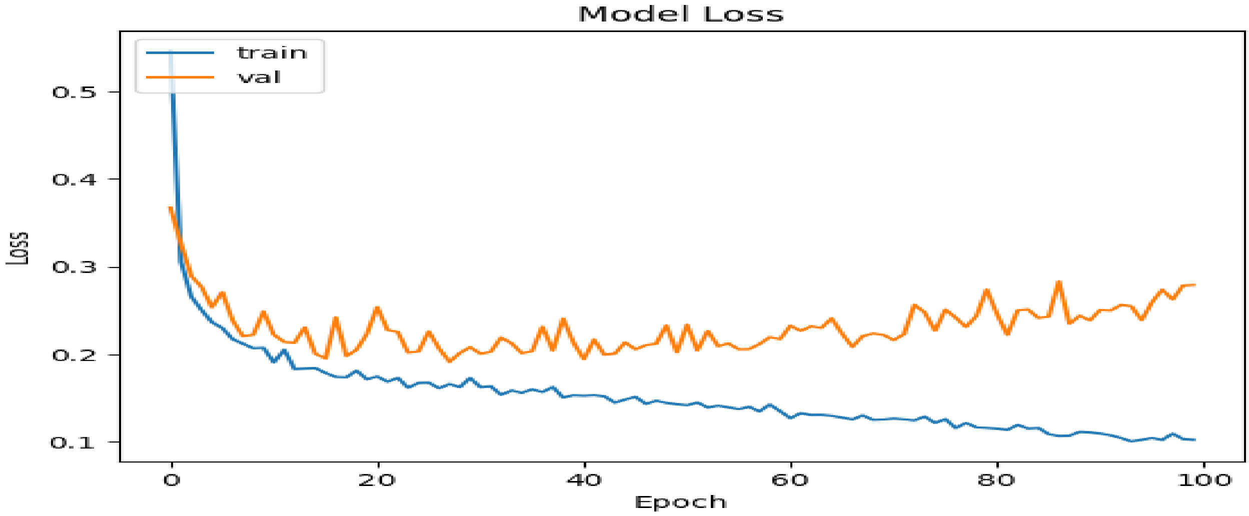
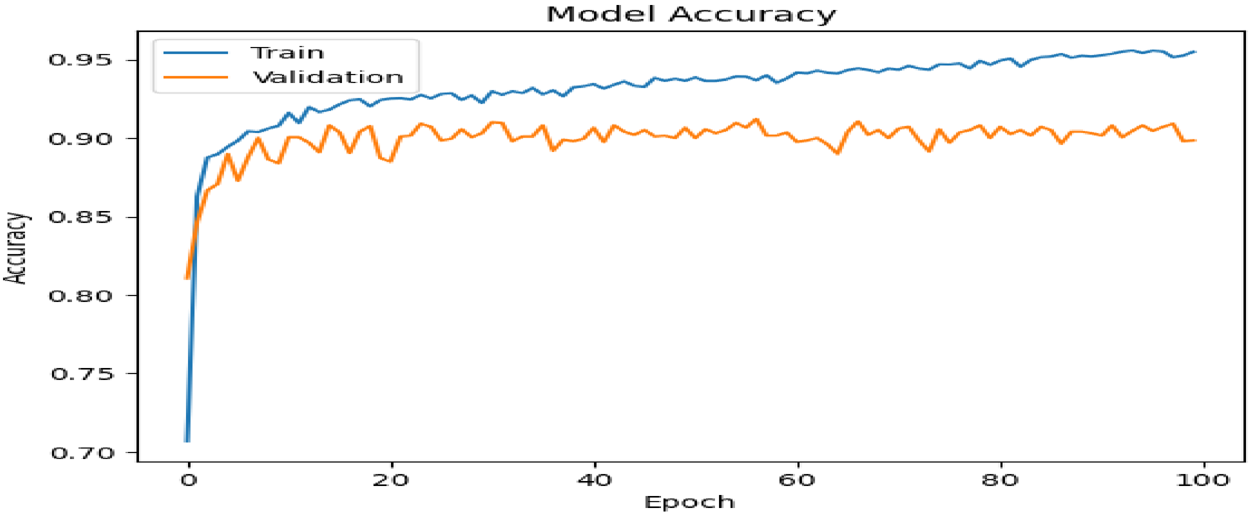


Figure 7: curve showing multilayermodel loss

5.1.3 Logistic regression

Logistic regression, standard scaler, classification report, and matplotlib are all available. A conventional scaler is used to scale the data, and logistic regression is trained on the training data. The model's accuracy is displayed for both the training and validation datasets.The classificationreport function is used to compute evaluation metrics such as precision, recall, and F1-score. Using matplotlib, the history of the previously trained deep learning model is shown for both training and validation accuracy and loss. Train accuracy:

0.5197, Validation accuracy: 0.5328 beginfigure[H]



curve showing Logistic regression accuracy

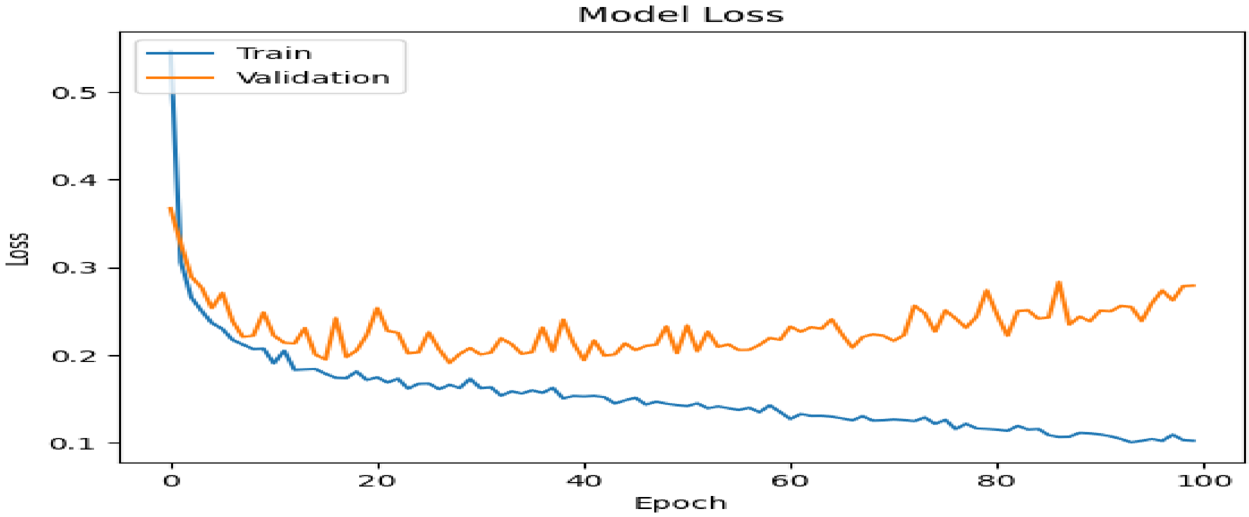


Figure 8: curve showing Logistic regression loss

5.1.4 Model Performance

Layers	Training Accuracy	Validation Accuracy
singlelayer	0.6167	0.6247
Multi layer	0.9614	0.9121
logical regression	0.5197	0.5328

Table 1: Performance comparison for different layers

6 Model Evaluation

When evaluating a classification model, three important metrics to consider are precision, recall, and f1-score. Precision measures the proportion of positive predictions that were correct, while recall measures the proportion of actual positive cases that were correctly identified. F1-score is an overall evaluation metric for classification models, where a value of 1 is the best and 0 is the worst.

Model	Precision	Recall	F1-Score
single layer	0.6276	0.8191	0.7107
multilayer	0.9277	0.9321	0.9299
logicalregression(+ve)	0.59	0.80	0.68
logicalregression(-ve)	0.20	0.08	0.12

The table shows the precision, recall, and F1-score for three different classification models. The single-layer neural network has a precision of 0.6276, recall of 0.8191, and an F1-score of 0.7107. The multilayer neural network has the highest precision of 0.9277, recall of 0.9321, and an F1-score of 0.9299. The logistic regression model for positive identification has a precision of 0.59, recall of 0.8, and an F1-score of 0.68. On the other hand, the logistic regression model for negative identification has the lowest precision of 0.2, recall of 0.08, and an F1-score of 0.12. Overall, the multilayer neural network model outperformed the other models in terms of all three evaluation metrics.

7 Feature Importance Analysis

As of now, we have an excellent grasp of the model to use, the amount of epochs to employ, and an acceptable validation set to use for the network design. The next stage is to determine which input attributes are redundant or unimportant.

7.1 Significance of individual features

The code above trains multiple neural network models to classify a dataset of features into a binary target variable. It starts by splitting the data into features (X) and target (y) and then trains models with single features, saving the best model for each feature. The validation accuracy of each single-feature model is stored in a list. Then, the code trains models with reduced feature sets, starting with the feature set that achieved the highest validation accuracy in the single-feature models. The code computes the validation accuracy for each reduced feature set and stores it in a list. Finally, the code plots the validation accuracies of both the single-feature models and the reduced-feature models. The plot shows that the validation accuracy generally decreases as features are removed from the model.

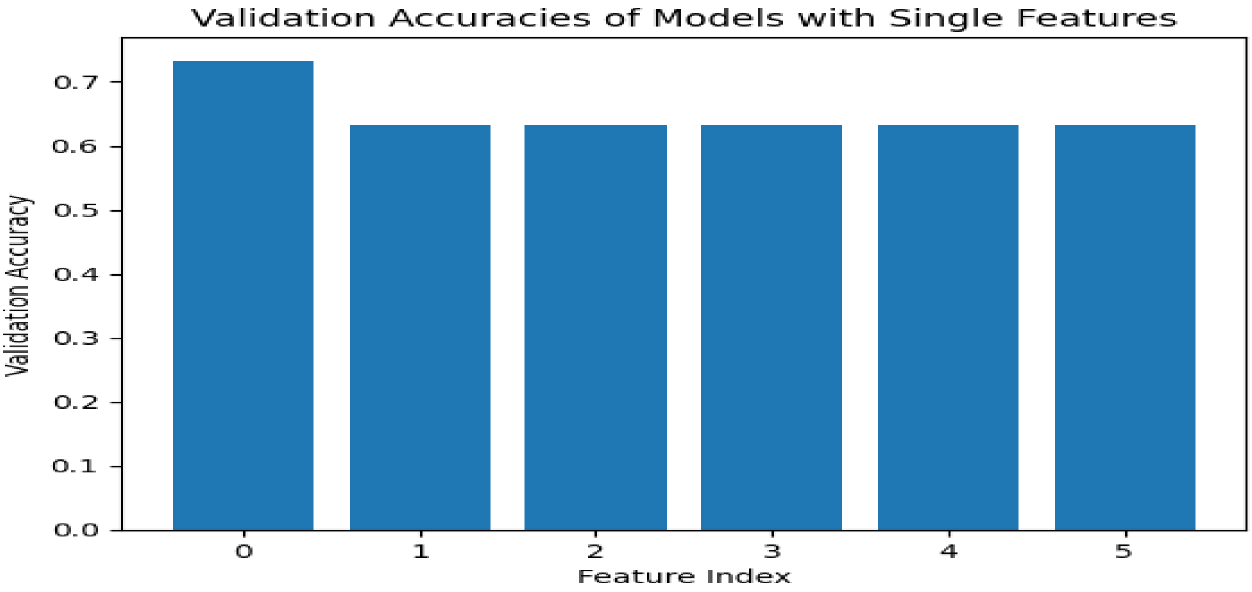


Figure 9: validation accuracys of model with single features

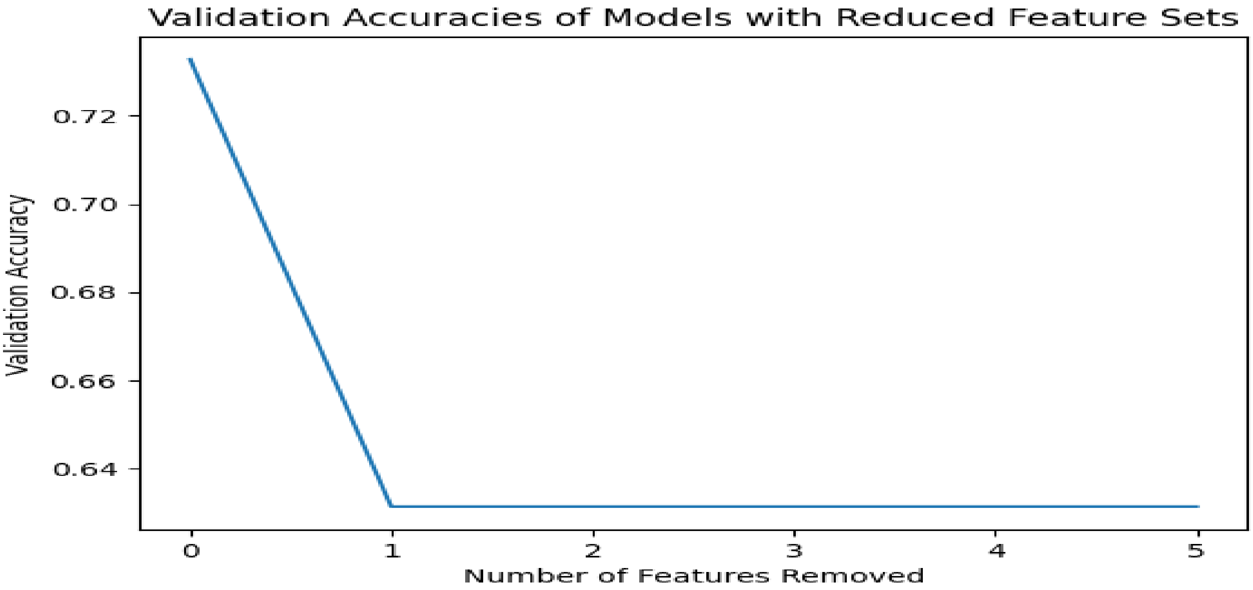


Figure 10: validation accuracys of model with single features

8 Challenges Faced

A number of variables, including respondent mistakes, interviewer errors, and data entry errors, may have an impact on the quality of the survey data collected. These may lead to erroneous or insufficient data. Consequently, obtaining excellent precision requires time. built a denser layer for better accuracy

9 Conclusion

To summarize, the motivation of the dataset can be helpful for market research, data analysis, and machine learning applications since it offers valuable information about customers' preferences for computer brand names. It makes it possible to comprehend how demographic characteristics affect consumer preferences for particular computer brands and can help in the development of prediction models for successful marketing tactics. In general, the dataset is a useful tool for companies looking to better understand their customers and develop goods and marketing strategies that cater to their demands.