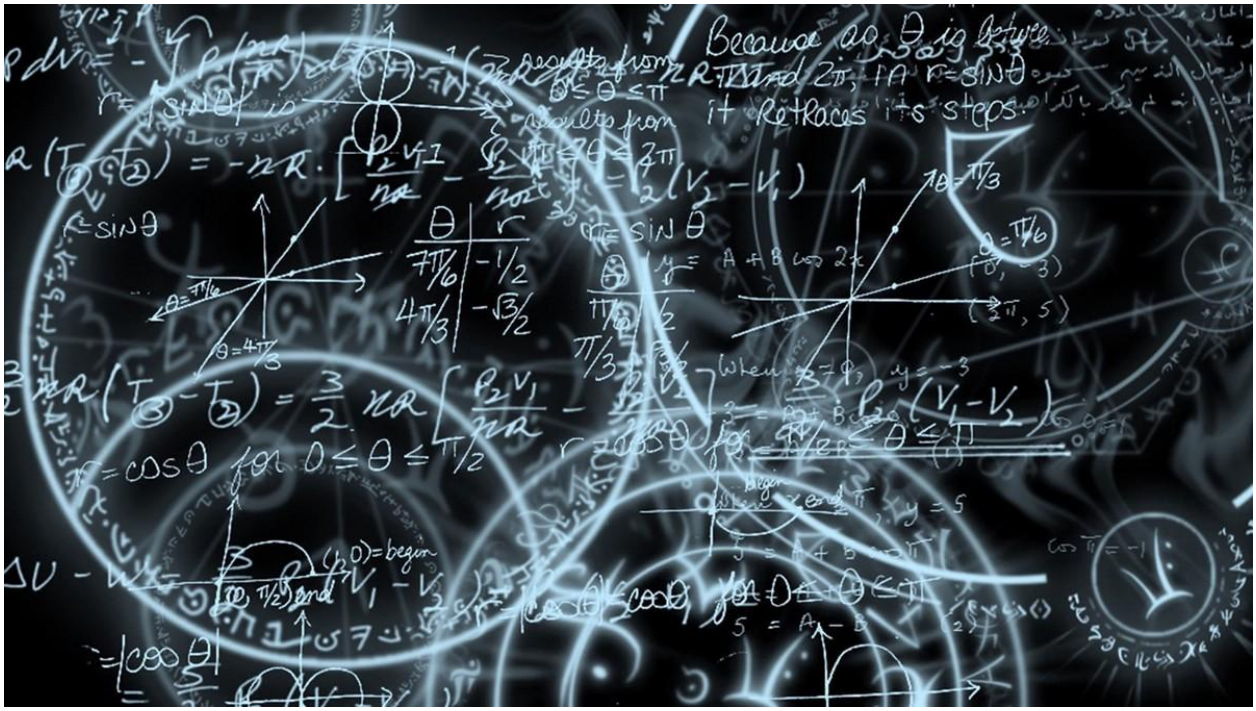
UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE



Module: Machine Learning (ST3139)

UOL Student No: 220640060

Number of Pages: 10 (Excluding the Cover Page, Table of Contents and Bibliography section)

Table of Contents

Summary Brief

This report details the implementation of three machine learning methodologies named **Unsupervised learning, Regression** and **Classification** on three distinct datasets, extracted from Kaggle. Each analysis was initiated with comprehensive EDA (Exploratory Data Analysis) to understand the data further and address appropriate research questions.

The links to each dataset will be available in the bibliography.

1.0 Unsupervised Learning

1.1 Introduction

It is a branch of machine learning where a model does not get trained under the presence of labeled data, hence requiring no human supervision to reveal new patterns or group similar data points to its own groups from unlabeled data. Some of the major applications include Anomaly detection, Recommendation systems, Medical Imaging and Computer Vision. (IBM, 2021)

The objective of this section will be to utilize dimensionality reduction and clustering to reveal customer behavior insights within a "credit card dataset", consisting of only 17 numerical behavioral variables such as the account balance, transaction frequencies, limits, payment details, etc. for approximately 9000 uniquely identified individuals during a 6-month timeframe, also noting that this is a static dataset (i.e., behaviors of a single cross section)

1.2 Existing Literature

No academic literature exists for this specific dataset, but with goals to reduce bad debts by limiting credit card capacities, a study on customer segmentation on this Kaggle dataset, (Karo, Yusmanto, & Setiawan, 2021) uses PCA and the K-means algorithm with the help of the silhouette index, to cluster 6 distinct groups consisting of installment payers and other shoppers. Given the static nature of this dataset, another study by (Nie, Chen, Zhang, & Guo, 2012) makes a longitudinal focus on a major Chinese commercial bank's credit card customers behavior, using distance-based techniques such as the DBSCAN algorithm. Outliers were eliminated at the 99th percentile while normalizing the data to reduce the effect from different dimensions.

1.3 Research Questions

- Can customers be segmented based on their spending habits and into how many groups if so?
- Can a group that potentially qualifies for an upgraded credit card status be identified?
- Can clusters be identified based on spending, repayment and cash advance levels?

1.4 Exploratory Data Analysis (EDA)

Data on figure 1.41 are evenly spread out between low and high purchase frequencies, indicating that some customers buy frequently and that the others buy rarely, but in similar proportions. But, the boxplot in figure 1.42 is **positively skewed** (i.e., most customers have lower credit limits, but a few have very high credit limits), making the distribution shift to the right.

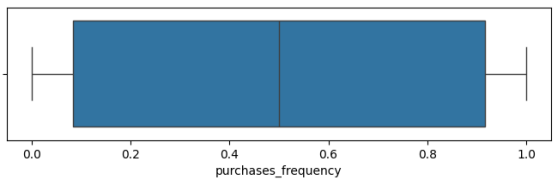


Figure 1.41

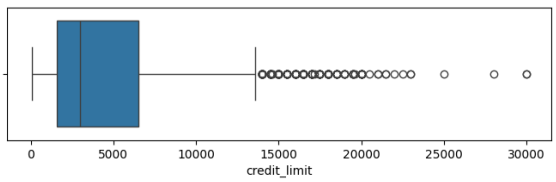


Figure 1.42

This indicates the presence of groups of customers exhibiting similar behaviors, hence making clustering a valid approach.

Figure 1.43 was plotted to identify and visualize **3 distinct customer groups** based on purchase frequency and credit limit available as shown on the right. The 3 groups identified were classified as customers with

01. **Low Credit, Low Spend** **–** Access to a smaller credit and spends rarely.

02. **High Credit, High Spend** **–** Access to a large credit limit and makes frequent purchases.

03. **High Credit, Low Spend** **–** Access to a large credit limit but rarely spends. (Inactive than expected)

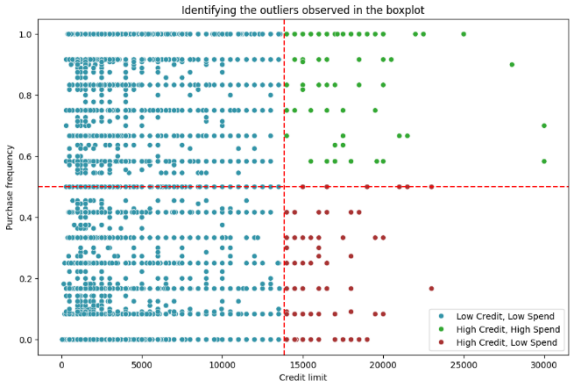


Figure 1.43 (Identification of potential groups)

Key Definition: Credit Card Utilization – A ratio that shows how much people use their credit limit relative to the available limit.

A new column, 'Credit Card Utilization', was constructed for further analysis.

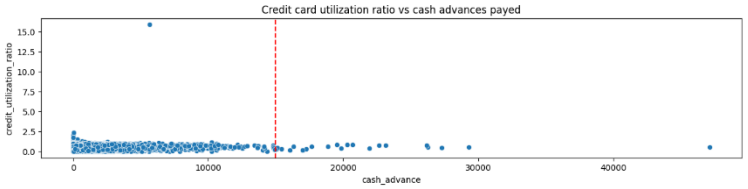


Figure 1.44

Figure 1.44 shows that most people take average cash advances. However, some take high cash advances in general (shown by the data points after the **red dotted line** at **15,000**).

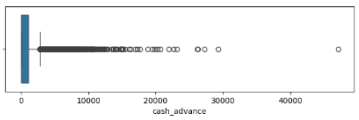


Figure 1.45

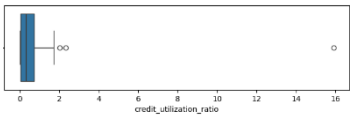


Figure 1.46

Figure 1.46 also highlights 3 outlier points, indicating high credit utilization ratios comparatively, but this is well below the threshold of 30% that presents as problematic, hence presents no issues.

1.5 Data Preprocessing

The dataset contained 313 missing values under 'minimum payments' and 1 missing value under 'credit limit'. To avoid data loss, missing values were replaced using an **Iterative imputer**. This looks for similar records or customers' information in the dataset to predict the missing numerical values. In order to make sure that it does not affect the distribution of the features after imputing, a histogram for the feature before and after imputing was plotted, where it showed no signs of changes.

Outliers (i.e., unusual values) were not removed in order to understand diverse customer behaviors, making sure to not miss on potential extreme cases that could help the bank make key decisions in terms of business perspectives.

At last, the features were scaled using a **Standard scaler** to ensure that features make equal contributions to the clustering process. This makes it comparable as if measurements were in the same units. And KMeans tend to perform better with Standard scaler as it relies on **Euclidean distance**. [Distance between 2 points] (KMeans, n.d.)

1.6 Principal Component Analysis (PCA)

PCA is a method used to simplify a dataset with many features into a smaller dataset (aka Dimensionality Reduction), while preserving maximum variance, hence making the model more manageable and the results much easier to interpret. The principal components formed will be a linear combination of the original variables and uncorrelated, while being ordered by the amount of variance they explain. (Whitfield, 2024)

A cumulative explained variance ratio graph was plotted to decide the optimal number of components required. As seen from the graph, the explained variance ratio reduces or diminishes with every extra component, hence the cutoff point was decided at 6 components which was able to explain 76% of the variance.

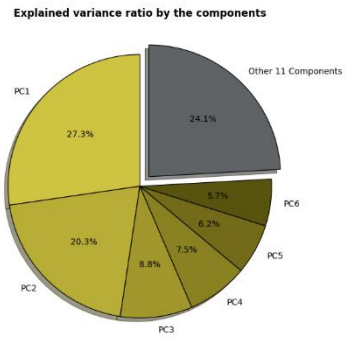


Figure 1.61

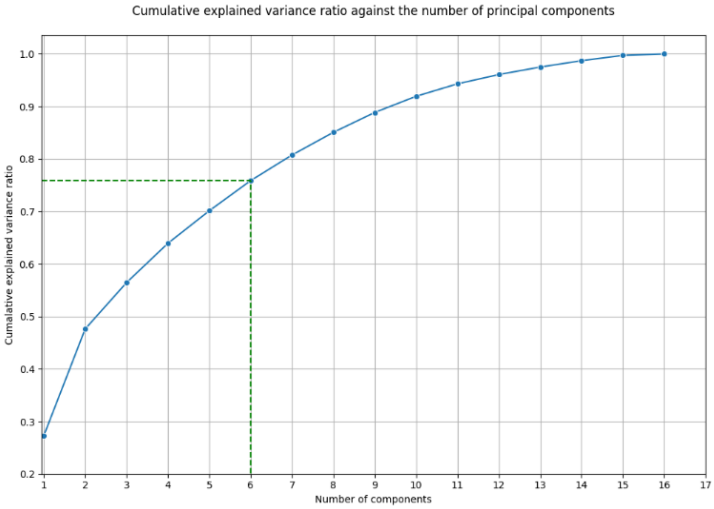


Figure 1.62 (Cutoff point at 6 components)

1.7 K-means Clustering

K-means is a clustering algorithm, where data points are assigned to k distinct clusters by minimizing the distance between data points and a cluster's center, aka centroid, while maximizing the distance between clusters to achieve well separated groups. (Sharma, 2025)

The optimal number of clusters, (k) was determined via an elbow plot. The number k decided will correspond to the place where the curve noticeably bends, (i.e., At **k=4** in this case) and figure 1.72 below shows the silhouette score. This shows how separable the clusters are, hence evaluating how successful the clustering process is. The silhouette score has a range between **-1** and **+1** and the higher being better. [**0.265** in this case] (Tomar, 2025)

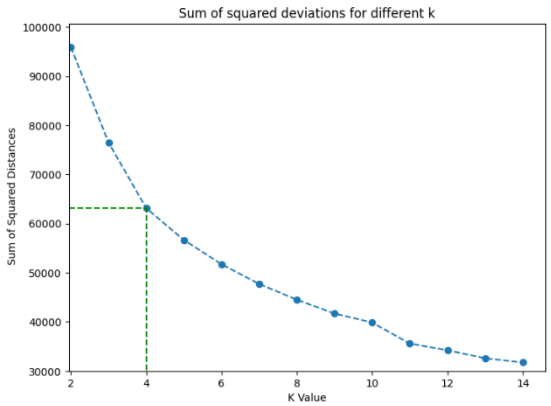


Figure 1.71

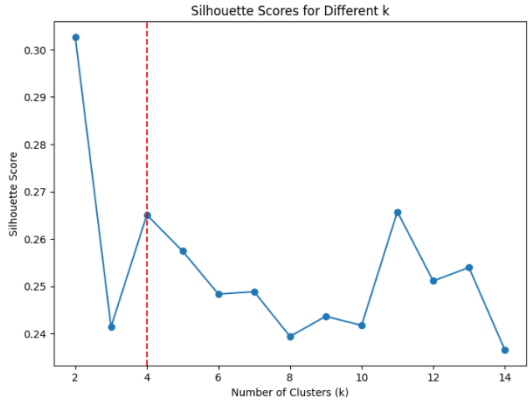


Figure 1.72

1.8 Findings and Interpretations

Data was grouped by each cluster to analyze differences in the mean values between each column before interpretations as shown by the data from the **cluster profile** below by table 1.81.

The figures 1.81 and 1.82 below were able to clearly show 4 clusters with only 2 PCA components and 3 PCA components respectively, which were covering majority of the explained variance ratio as was demonstrated by the pie chart in figure 1.61.

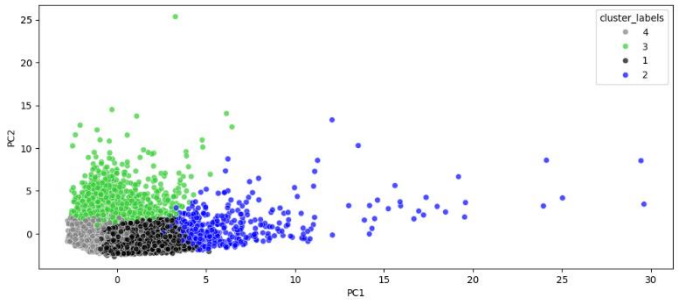


Figure 1.82 (Cluster separation for 2 PCA components)

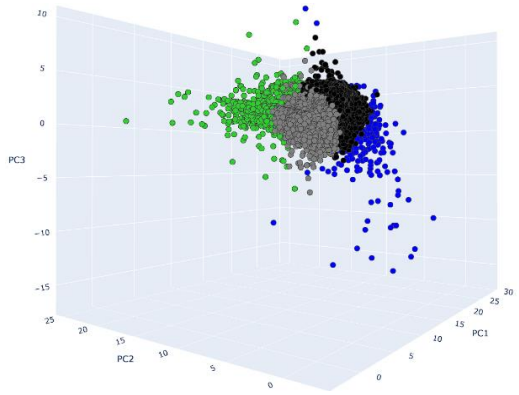


Figure 1.83 (Cluster separation for 3 PCA)

Cluster 1: Represents the frequent customers, that tends to pay in installments. Maintains a low balance overall and pays in full around 27% of the time. **(Ideal behavior for an upgraded status**.)

Cluster 2: Represents the high spending customers with a higher tendency to make one off payments. Their transactions are backed by the highest credit limit on average and also pays in full around 29% of the time.

Cluster 3: Represents customers that withdraw cash from the credit card, showing less dependency on credit cards to make purchases. Carries significant debt, as can be seen by the high balances and the low tendency of around 4% to pay in full ,given its for cheaper purchases.

Cluster 4: Represents customers with minimal credit card usage or activity. The lowest credit limits and inactivity can be seen from most of the rows.

	cluster_labels	1	2	3	4
balance	mean	892.832791	3532.141485	4512.579181	1006.905485
balance_frequency	mean	0.940730	0.985436	0.968393	0.786661
purchases	mean	1236.290684	7609.855664	476.479307	282.149697
oneoff_purchases	mean	580.526098	5106.647464	299.400832	218.461628
installments_purchases	mean	656.046676	2504.630000	177.161403	63.951579
cash_advance	mean	209.516159	691.341086	4477.685023	569.313053
purchases_frequency	mean	0.890692	0.945225	0.280309	0.181400
oneoff_purchases_frequency	mean	0.297179	0.742647	0.134747	0.089534
purchases_installments_frequency	mean	0.719847	0.781161	0.180355	0.088084
cash_advance_frequency	mean	0.042699	0.073241	0.483394	0.110757
cash_advance_trx	mean	0.796517	2.132701	14.203915	2.024324
purchases_trx	mean	22.544455	86.232227	7.429853	3.069248
credit_limit	mean	4177.442790	9765.521327	7411.953137	3311.805693
payments	mean	1322.897592	7271.084160	3463.578003	959.803800
minimum_payments	mean	636.252653	1954.485313	1951.510546	589.418711
prc_full_payment	mean	0.270880	0.289506	0.036279	0.080046
tenure	mean	11.628781	11.938389	11.407015	11.416232

Table 1.81 (Cluster profile)

More groups with detailed justifications were identified through K Means clustering compared to the basic groupings assumed by figure 1.43. However, this is lesser, compared to the results from existing literature.

Supervised Learning

Uses datasets that have a labeled target to train models into identifying patterns or relationships between inputs and outputs, making it a goal to generalize the model to accurately predict on unseen data. (IBM, 2021) Datasets for the latter tasks were trained on 80% of the data and the remaining 20% to evaluate each model's results. Data preprocessing was done separately for each set to avoid any data leakage, so that no answers or patterns from training slip out until evaluation.

2.0 Regression

2.1 Introduction

Utilized for exploring relationships between independent variables aka features and a dependent variable to predict continuous outcomes after getting trained on labeled data. (Machnie Learning Regression Explained, 2021)

This section focuses on predicting the "Medical Insurance cost" based on 6 features and 1337 unique records. The features include age, gender, number of children covered by health insurance, region of stay with US, whether the individual is a smoker or not and the BMI (Body Mass Index).

2.2 Existing Literature

(Billa, 2024) uses a larger dataset containing 2773 records, but with the exact same structure and features to understand the potential of Machine Learning on predicting Medical Insurance costs. The best performing model was a Gradient Boosting Regression with a R^2 value of 0.8679, MAE of 2389.9 and a RMSE of 4453.8. This article also highlights additional sources such as satellite imagery and social sentiment analysis's ability to improve the model's predictive power.

2.3 Research Questions

1. Which factors are the most influential on medical cost charges?
2. Does smoking have an effect on medical cost charges and does this change by region?
3. Can we identify the combination of factors that would produce the highest charges and the lowest charges?

2.4 Exploratory Data Analysis (EDA)

A heatmap was plotted during EDA to check for any correlations (strength of the linear relationship) between numerical variables. Weak correlations of 0.3 were observed between age and charges, 0.2 between age and charges at best.

It can be seen from the line plot on figure 2.41, that charges increase overall with increasing age as expected. The fluctuations indicate that there are certainly other factors affecting the charges.

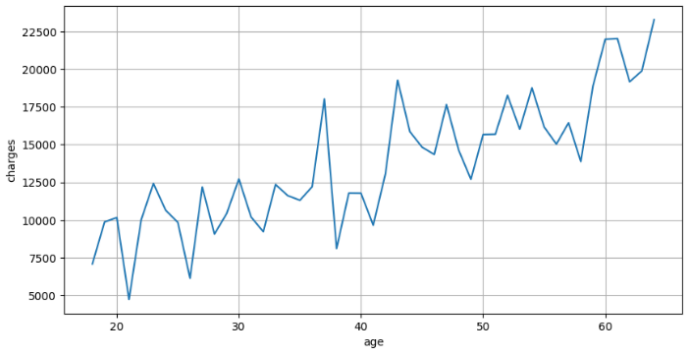


Figure 2.41 (Charges against age)

Calculations showed that the ratio of average charges incurred between smokers to non-smokers are approximately 3.8 times. Hence, the next few visualizations were constructed to go in depth and check for any other factors.

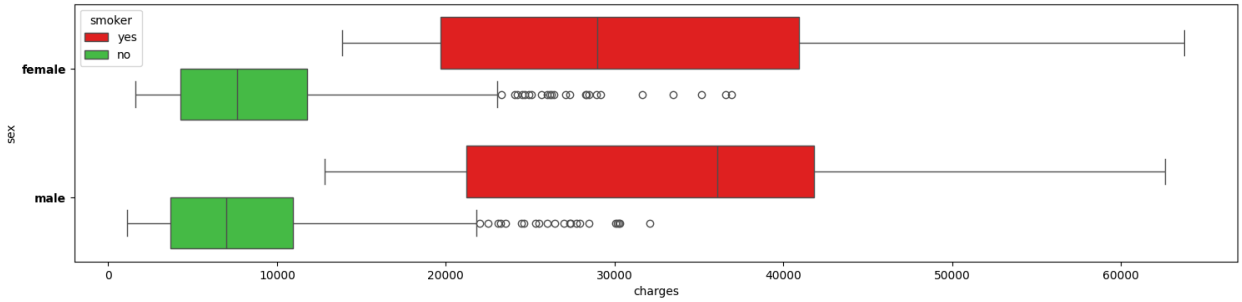


Figure 2.42 (How charges vary depending on whether a person smokes and the Gender)

The median charges were split by gender and observed on a box plot. It is very evident on how much the charges increase when people smoke. It can also be seen that the median charges are high for individuals who are male and smoking.

The grouped bar chart on the next page in figure 2.43 clearly shows that people who smoke incur higher charges on average in the **southwest** region. Further calculations show that 1 in 4 people in the southeast region smoke, which is the highest in any region, and this suggests that the southeast region might be charging higher due to increased risk assessments.

However, non-smokers tend to incur higher charges on average in the **northeast** region, but the differences in average charges amongst non-smokers are not significant.

This suggests that different regions might be calculating medical insurance costs at different rates based on whether a person smokes or not. Other possible factors could include cost of living, health policies, mean age of the population, possible diseases in particular regions, etc.

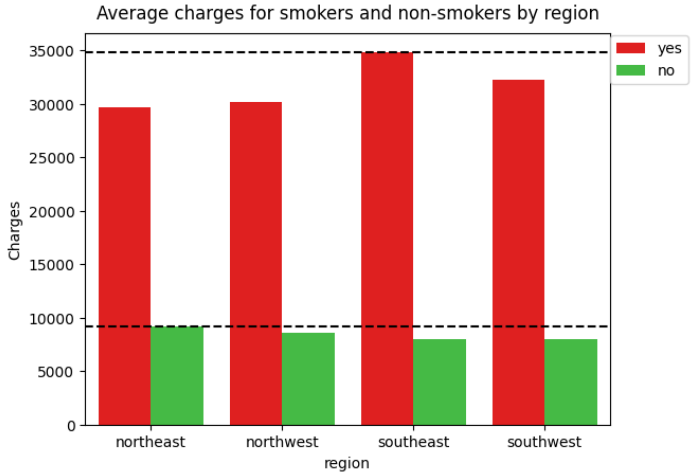


Figure 2.43 (Grouped bar chart)

Figures 2.44 and 2.45 show how BMI and being a smoker, **jointly** amplifies the charges, as shown by the trend of the red points. BMI was broken down into 5 categories, showing clear groupings.

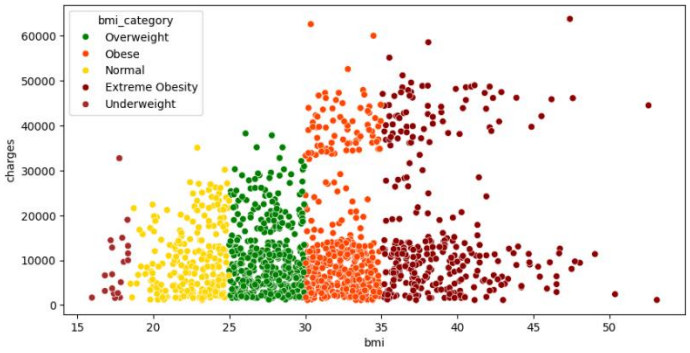


Figure 2.44 (BMI vs Charges separated by bmi-category)

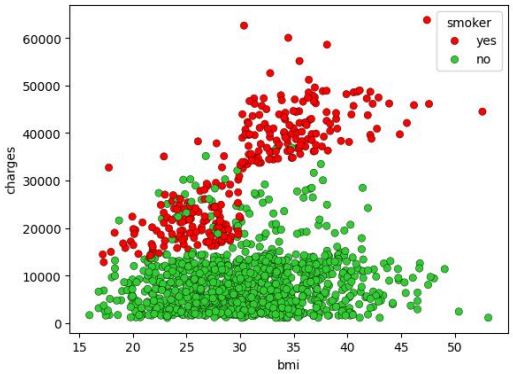


Figure 2.45(BMI vs Charges based on smoking)

2.5 Data Preprocessing

The dataset had no missing values, but contained potential outliers in the 'bmi' and the 'charges' features, hence the interquartile range was used to remove the outliers but with a **multiplier set at 3** instead of the standard 1.5, allowing more flexibility to accept outliers, addressing the unpredictable nature of this problem. Any points beyond the range were removed.

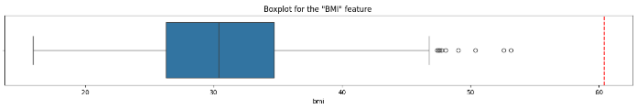


Figure 2.51

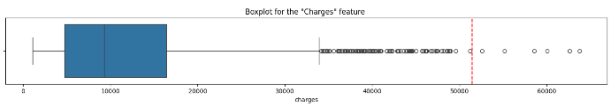


Figure 2.52

The **nominal categorical** features, 'smoker' and 'region' were encoded, specifically one hot encoded, converting categorical data into numerical data so that ML algorithms can understand. (Kristianto, 2023)

Key Definition: Nominal – Categories or labels in a feature containing categorical data have no order or rank.

And at last, the features were scaled using a Standard scaler.

Key Definition: Scaling – A process to improve model performance, reduce the impact of outliers and ensure that data is on the same scale. (Bhandari, 2025)

2.6 Feature Selection

During this stage, the goal is to select features that have high correlations with the target variable, 'charges' and remove any features that have high correlations amongst other features to reduce any redundancy in data, aka multicollinearity.

After inspecting the heatmap, the features, 'region_southeast' and 'region_northwest' were removed, as it showed poor correlations with the target variable (i.e., charges). However, one of the encoded variables, 'region_southwest' was not removed due to the relationships revealed during EDA in connection to region and also to consider some data input from region without eliminating all information originating from region. The feature, 'smoker' was the most correlated with the target variable, 'charges' with a value of 0.79.

And in this heatmap, correlations were visualized for the training data after encoding and before running any models. This shows the features, 'smoker_yes' to have a strong positive correlation of 0.79 and the feature, 'age' to have a correlation of 0.3.

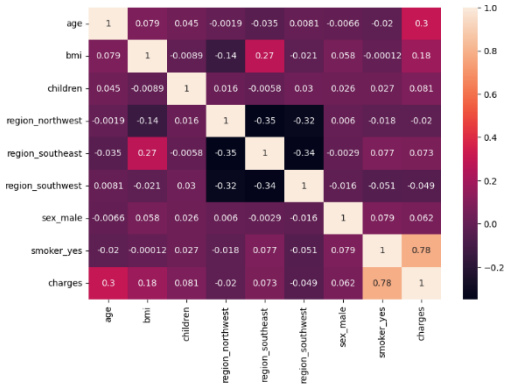


Figure 2.61 (Heatmap showing correlations on the train data)

2.7 Regression Models

After hyper tuning the parameters of each model, the main goal was to minimize the 3 **loss functions** elaborated below while maintaining a high R-squared value (Between 0 and 1).

Key definitions

- Residual – The difference or the deviation between the actual and the predicted values.
- Mean Absolute Error (**MAE**) - Average of the absolute deviations between the between the actual and the predicted values.
- Mean Squared Error (**MSE**) - Average of the squared deviations between the actual and the predicted values.
- Root Mean Squared Error (**RMSE**) - Square root of the MSE, aka the standard deviation of the residuals.
- R-squared value(**R^2**) - Measures the variance of the actual values that the model can explain.

Model	R^2	MAE	RMSE	MSE
Multiple Linear Regression	0.7835	4119.99	5719.80	3.272×10^7
K-Nearest Neighbors (KNN)	0.8414	3230.57	4896.04	2.397×10^7
Random Forest Regressor	0.8934	2330.51	4013.91	1.611×10^7
Gradient Boosting Regressor	0.8944	2286.87	3994.07	1.595×10^7
XG Boost Regressor	0.8934	2354.46	4012.85	1.610×10^7

2.8 Results and Findings

5 models were tested, and tuned for optimal hyperparameters. The results, after evaluating on the test set were compared to identify the best performing model in predicting medical insurance costs. 3 models performed significantly well, being the Random Forest, XG Boost and the Gradient Boosting Regressors, where the Gradient Boosting Regressor as highlighted above in the table, gained the slight edge in achieving the highest R^2 value and in minimizing the loss functions that were discussed above. This aligned with results from past studies and even performed better, but should be noted that the past study was conducted on a larger dataset. This could make our model perform better or worse, so must be evaluated to confirm.

Multiple Linear Regression was the worst performing model, despite the R^2 value of 0.7836. The error terms were **not** normally distributed and residual plots revealed uncaptured patterns, **violating assumptions** and indicating **non-linear** behavior that cannot be captured by a linear regression model.

Furthermore, the most successful model, (i.e., the Gradient Boosting regressor) was used to assess feature importance. The model has mainly depended on the features, **'smoker_yes'**, followed by **'bmi'**, **'age'** and **'children'** in the magnitudes shown in figure 2.81, when making predictions.

The features, 'region_southwest' and 'sex_male', in other words no information provided by **region** and **gender** wasn't found useful by the model.

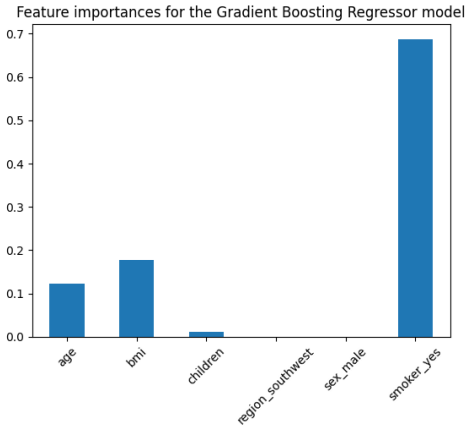


Figure 2.81 (Feature Importances)

3.0 Classification

3.1 Introduction

Classification assigns a predefined label of a target variable on unseen input data. The model uses each observation's features against its class label to understand which features define each class. (IBM, 2021)

The following dataset contained a total of 5842 reviews in which 1852 were positive, 860 were negative and 3130 were neutral. Sentiments with the label, 'neutral' were removed for this task, where reasonings will be discussed in section 3.6

This task focuses on a binary classification problem, specifically on a sentimental analysis in the financial domain. The goal will be to correctly classify the sentiment (positive, negative) for a particular sentence within the corpus of financial texts, in turn allowing to predict market fluctuations and aid investment decisions in the corporate world.

3.2 Existing Literature

(Adelakun & Baale, 2024) had evaluated their models on this exact same dataset. The following methodology had down sampled the reviews to the count of negative reviews. (i.e., 860 reviews of each type). This study has prioritized deep learning using a BERT model (Bidirectional Encoder Representations from Transformers) to achieve extremely high results of 95.24% accuracy, and a F1-score of 95.32%.

3.3 Research Questions

- Can we accurately classify between positive and negative sentiments given a particular review?
- Can we identify specific keywords or phrases that strongly correlate with particular sentiment labels?

3.4 Exploratory Data Analysis (EDA)

The word cloud in figure 3.41 displays the most common words contained in sentences with the positive sentiment and for the negative sentiment in figure 3.42. The size of each word corresponds to the number of times it repeats.

In the positive sentiment word cloud, familiar words can be seen such as **'profit', 'sale', 'increase', 'new'** and **'higher'**, **'long'** etc. reflecting on the positivity and any improvements in financial performance. Similarly, common words in the negative sentiment word cloud includes **'lower', 'sale', 'decrease', 'loss'**, **'fall',** etc.

However, similar words can be found in both word clouds such as **'sale', 'profit, 'net profit', 'operating profit',** etc. showcasing the importance of context.

*Figure 3.41 (Word cloud for the **positive** sentiment)*

*Figure 3.42 (Word cloud for the **negative** sentiment)*

The **average number of words when training data per sentence** was 12 when used for machine learning models and 20 when used for deep learning models. This was because the stop words were not removed for DL models as mentioned before.

3.5 Data Preprocessing

Since the dataset contains manually typed reviews, it is prone to a lot of errors such as **punctuation, whitespace, misspellings, meaningless words, special characters, links, etc.** that do not improve a model's predictive power significantly, hence were removed to reduce noise.

For **ML models**, stop words (i.e., words like the, is, and more) were removed as these words do not improve model performance, but was removed for **DL models**, since it could be problematic for deep learning models if sentences lose their semantic meaning as they can learn the context better.

Machines understand numbers only, hence textual data needs to be converted to numerical data and this process is called vectorization. In this case, **TF-IDF** vectorization was used for the ML models. A 'Term Frequency Inverse Document Frequency' Vectorizer is used to store how significant or important each input word is.

In the process of vectorization, the words first get split into individual words (aka tokens) and then get converted to numbers. But before applying a DL model, the resulting tokens get additionally passed on to an embedding layer. In this case word embeddings were used, out of several other methods such as including character embeddings, sentence embeddings, etc.

Embeddings form dense vectors that can have any dimensions and need not get padded with 0s like in vectorization to create sparse matrices, making it more suitable for downstream tasks and is better in representing meaning, even though computationally expensive and this can get adjusted as the model learns rather than staying fixed like in vectorization.

After removing the neutral sentiments, the counts of each label were as displayed in figure 3.51 mentioned below.

Imbalance was seen in the data. This happens when one class oversaturates the labeling, causing the model to favor that dominant class during predictions. Even though existing literature down sampled the data, over sampling using SMOTE was the technique chosen to avoid data loss

This technique creates synthetic samples of existing data to make the count of negative sentiments in the train data to equal the counts of the positive reviews (i.e., 1862)

This ensures that the model has no bias during predictions.

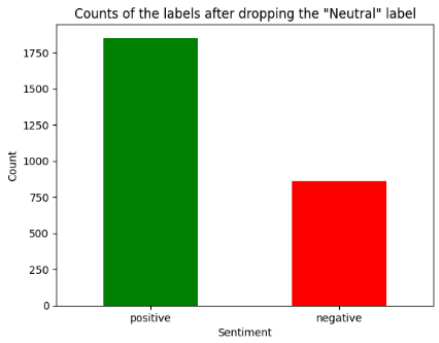


Figure 3.51 (Label counts)

3.6 Classification approach and justification

The naïve bayes algorithm is one of the most popular models in NLP (Natural Language Processing) applications. Hence was fitted on the training data to workout 2 implementations in order to decide whether to proceed with the 3 classes (i.e., as a multi-classification problem) or with a binary classification problem with 2 labels only.

The 2nd implementation was chosen as the model performed significantly better and showed more reliable results. A binary classification problem is much easier to interpret and understand. And another decision to remove the 'neutral' label was to simplify the analysis and focus more on the sentiment polarity (Positive or Negative). And in the 1st implementation, there would have been too much of data loss or additional synthetic data after adjusting to fix imbalance issues.

3.7 Model Building

3 machine learning (ML) models, 4 deep learning (DL) models containing a manually built neutral network architecture and a few using pretrained models will be implemented.

Key Definition: Pretrained model – These are models that get trained on large datasets in a particular field. The knowledge the model gains get stored and saved in weights (Numbers that show how much each input matters). These can be used for similar problems avoiding or minimizing the need to decide on architectures when building neural networks.

The concept of **transfer learning** was implemented on the last 3 models mentioned in the table below during training. It is a common application in deep learning projects which uses pretrained models for training, where the learned weights from models are used, making it convenient and time efficient. (Gupta, 2025)

2 pretrained models were used, being **USE** (Universal Sentence Encoder) and **BERT** (Bidirectional Encoder Representations from Transformers). These were used due to their popularity in tasks where context is key. Financial context, in this case.

The models can be set to non-trainable, where the exact fixed weights get used to evaluate and make predictions for the problem or non-trainable, where the weights get fine-tuned or adjusted during training to better fit to our model.

- True positive (**TP**) – The model predicts positive and its correct.
- False positive (**FP**)– The model predicts positive but its incorrect.
- True negative (**TN**) – The model predicts negative and its correct.
- False negative (**FN**) – The model predicts negative but its incorrect
- he following metrics were used to evaluate the models mentioned above.

(Kumar, 2025) states the following metrics that will be prioritized to evaluate the models.

Metric	Definition	Calculation
Accuracy	Measures how often the model correctly classifies.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	The proportion of correctly predicted cases that are actually positive.	$\frac{TP}{TP + FP}$
Recall	Checks how many of the actual positive values were plotted correctly.	$\frac{TP}{TP + FN}$
F1 Score	A combined idea of precision and recall.	$2 \times \frac{Precision \times Recall}{Precision + Recall}$

3.8 Results and Findings

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.80	0.79	0.80	0.79
Logistic Regression	0.80	0.79	0.80	0.79
Gradient Boosting Classifier	0.79	0.79	0.79	0.79
Simple Neural Network	0.83	0.82	0.83	0.82
USE model: Non-trainable	0.78	0.77	0.78	0.77
BERT model: Non-trainable	0.76	0.77	0.76	0.76
USE model: Trainable	0.85	0.85	0.85	0.85

Accuracy wouldn't be as reliable due to the imbalance noted in the beginning, even though the labels were made equal using over sampling. Hence the F1- Score will be mainly used to identify the best model.

And clearly, the USE pretrained trainable model, performed the best while other models gave significant results.

The confusion matrix of the best performing model was used to take a closer look. As demonstrated by the figure, 39 negative sentiments were falsely predicted as positive and 45 positive sentiments were predicted as negative, and these 2 values should improved to better assist decision making in this context.

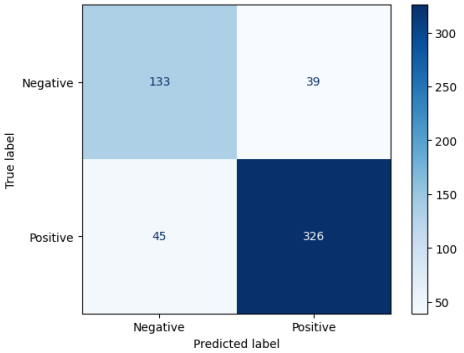


Figure 3.81 (Confusion Matrix)

Bibliography

(2021, September 23). Retrieved from IBM: https://www.ibm.com/think/topics/unsupervised-learning

Adelakun, N. O., & Baale, A. A. (2024, July 18). *SENTIMENT ANALYSIS OF FINANCIAL NEWS USING THE BERT MODEL*. Retrieved from ResearchGate: https://www.researchgate.net/publication/382309719_SENTIMENT_ANALYSIS_OF_FINANCIAL_NEWS_USING_THE_BERT_MODEL

Bhandari, A. (2025, March 10). *Feature Scaling: Engineering, Normalization, and Standardization (Updated 2025)*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Bhasin, A. (2018). *Credit Card Dataset for Clustering*. Retrieved from Kaggle: https://www.kaggle.com/datasets/arjunbhasin2013/ccdata/data

Billa, M. M. (2024, May). *Medical Insurance Price Prediction Using Machine Learning*. Retrieved from ResearchGate: https://www.researchgate.net/publication/381277259_Medical_Insurance_Price_Prediction_Using_Machine_Learning

Cho, M. (2018). *Medical Cost Personal Datasets*. Retrieved from Kaggle: https://www.kaggle.com/datasets/mirichoi0218/insurance

Gupta, D. (2025, Jan 30). *Transfer Learning Using Pre-trained Models in Deep Learning*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/transfer-learning-the-art-of-fine-tuning-a-pre-trained-model/#h-what-is-transfer-learning

Karo, I., Yusmanto, A., & Setiawan, R. (2021, November 17). *Segmentation of Credit Card Customers Based on Their Credit Card Usage Behavior using the K-Means Algorithm*. Retrieved from ResearchGate: https://www.researchgate.net/publication/374032027_Segmentation_of_Credit_Card_Customers_Based_on_Their_Credit_Card_Usage_Behavior_using_The_K-Means_Algorithm

KMeans. (n.d.). Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Kristianto, N. G. (2023, June 18). *Decoding the Power of Encoding in Machine Learning*. Retrieved from Medium: https://medium.com/@nicholasgabrielkr/decoding-the-power-of-encoding-in-machine-learning-39572e9cc6a3

Kumar, S. (2025, April 01). *Evaluation Metrics For Classification Model*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

Machnie Learning Regression Explained. (2021, October 29). Retrieved from Seldon: https://www.seldon.io/machine-learning-regression-explained/

Nie, G., Chen, Y., Zhang, L., & Guo, Y. (2012). *Credit card customer analysis based on panel data clustering*. Retrieved from ResearchGate: https://www.researchgate.net/publication/220307717_Credit_card_customer_analysis_based_on_panel_data_clustering

Sbhatti. (2022). *Financial Sentiment Analysis*. Retrieved from Kaggle: https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis

Sharma, P. (2025, March 21). *K-Means Clustering Algorithm*. Retrieved from Analytics Vindya: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Tomar, A. (2025, March 13). *Elbow Method in K-Means Clustering: Definition, Drawbacks, vs. Silhouette Score*. Retrieved from built in: https://builtin.com/data-science/elbow-method

Whitfield, B. (2024, Februrary 23). *Principal Component Analysis*. Retrieved from built in : https://builtin.com/data-science/step-step-explanation-principal-component-analysis