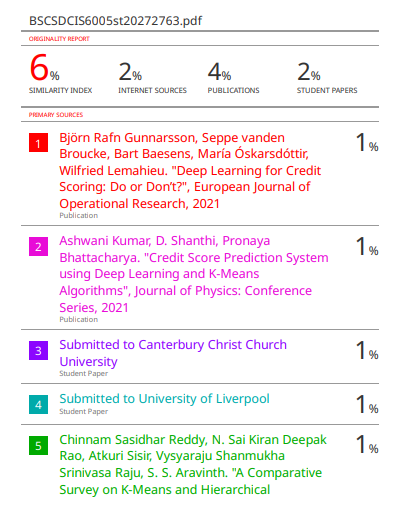
**Cardiff Metropolitan University**

**Assignment Cover Sheet**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student Details (Student should fill the content)** | | | | | | | | | | |
| Name | | | Sharmila Sithravelayutham | | | | | | | |
| Student ID | | | JF/BSCSD/12/06 (20272763) | | | | | | | |
| **Scheduled unit details** | | | | | | | | | | |
| Unit code | | |  | | | | | | | |
| Unit title | | |  | | | | | | | |
| Unit enrolment details | | | Year | | 3 | | | | | |
| Study period | | 2022-2023 | | | | | |
| Lecturer | | |  | | | | | | | |
| Mode of delivery | | | Full Time | | | | | | | |
| **Assignment Details** | | | | | | | | | | |
| Nature of the Assessment | | |  | | | | | | | |
| Topic of the Case Study | | | Deep learning for Finance plus AI mini project | | | | | | | |
| Learning Outcomes covered | | | L01, L02, LO3 | | | | | | | |
| Word count | | | 4000 | | | | | | | |
| Due date / Time | | | 05th September 2023 | | | | | | | |
| Extension granted? | | | Yes | No | Extension Date | | | |  | |
| Is this a resubmission? | | | Yes | No | Resubmission Date | | | |  | |
| **Declaration** | | | | | | | | | | |
| I certify that the attached material is my original work. No other person’s work or ideas have been used without acknowledgement. Except where I have clearly stated that I have used some of this material elsewhere, I have not presented it for examination / assessment in any other course or unit at this or any other institution | | | | | | | | | | |
| Name/Signature | | | Sharmila Sithravelayutham | | | | Date | | 2023.09.04 | |
| **Submission** | | | | | | | | | | |
| Return to: | | |  | | | | | | | |
| **Result** | | | | | | | | | | |
| Marks by 1st Assessor |  | Signature of the 1st Assessor | | | | | |  | | **Agreed Mark** |
| Marks by2nd Assessor |  | Signature of the 2nd Assessor | | | | | |  | |
| **Comments on the Agreed Mark.** | | | | | | | | | | |
| **For Office use only (hard copy assignments)** | | | | | | | | | | |
| Receipt date |  | | Received by | | |  | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STUDENT NAME:** | | | | | | | | | | **STUDENT NUMBER:** | | |
| **Module Number & Title**: | | | | | | | | | | **Semester:** | | |
| **Assignment Type & Title:** | | | | | | | | | | | | |
| **For student use: *Critical feedback on the individual progression towards achieving the assignment outcomes*** | | | | | | | | | | | | |
|  | | | | | | | | | | | | |
| **For the Assessors’ feedback**  **Indicate the Task number strength and Weaknesses and the marks for each task** | | | | | | | | | | | | |
| **Task No/Question No**  **CMU B.Sc. (HONS) SE- ASSIGNMENT FEEDBACK SHEET –ICBT CAMPUS** | **Strengths (1st Assessor)** | | | | | | **Strengths (2nd Assessor)** | | | | | |
| **Task No / Question No** | **Weaknesses (1st Assessor)** | | | | | | **Weaknesses (2nd Assessor)** | | | | | |
| **Areas for future improvement** | | | | | | | | | | | | |
| **Comments by 1st Assessor** | | | | | | **Comments by 2nd Assessor** | | | | | | |
| **Marks** | | | | | | | | | | | | |
| **Task /Question No** | | **Marks by 1st Assessor** | **Marks by 2nd Assessor** | | **Marks by IV (if any)** | | | **IV comments (If Any)** | | | | |
|  | |  |  | |  | | |  | | | | |
|  | |  |  | |  | | |
|  | |  |  | |  | | |
| **Total Marks** | |  |  | |  | | |
| **Name and the Signature of the 1st Assessor** | | | |  | | | | | **Date:** | |  | |
| **Name & Signature of the 2ndAssessor :** | | | |  | | | | | **Date :** | |  | |
| **Name & Signature of the IV: (If any)** | | | |  | | | | | **Date:** | | |  |



**Table of Contents**

**Applying Deep Learning for Credit scoring. | Task 01**

1. Acknowledgment ............................................................... ................................................ 2

2. Abstract ............................................................................................................................... 3

3. Introduction ........................................................................................................................4

4. Literature Review ................................................................................................................4-5

5. Discussion ...........................................................................................................................6

6. Conclusion .......................................................................................................................... .7

7. References ........................................................................................................................... 8

**Mall Customer Segmentation Analysis | Task 02**

Background............................................................................................................................... 9

The problem..................................................................................................................1:1

Literature Review ........................................................................................................1:2

Similar applications include..........................................................................................1:3.

How does this application differ ………………………………............................................. 10-11

What is the machine learning technique used?......................................................................... 12-16

Theory behind the machine learning ……………………...................................................... 12-16

Exploratory Data Analysis……………………...................................................................... 18-21

Conclusion ……………………................................................................................................21

References……………………..................................................................................................22

**Acknowledgement**

I want to extend my heartfelt appreciation to my lecturer, Ms. Laneesha, for the invaluable guidance and unwavering support she provided throughout the creation of this system report. Ms. Laneesha's profound insights and thoughtful suggestions have significantly enriched my comprehension of the system design and development process.

In addition, I wish to convey my gratitude to my classmates and friends who generously shared their insights and provided constructive feedback. Their contributions have played a pivotal role in shaping the final version of this report.

Lastly, I would like to express my deep gratitude to the authors of the research papers and books referenced in this report. Their work has served as an invaluable wellspring of knowledge, enriching my understanding of the pertinent concepts and principles.

**With sincere appreciation,**

**Sharmila Sithravelayutham.**

**Applying Deep Learning for Credit scoring.**

**Abstract**

In this study, we present a score for credit prediction system for the financial industry that uses deep learning in conjunction with the K-Means method. The system includes a prediction model that combines feature selection (FS) classification with deep learning applications, allowing for successful model training.

While numerous classification algorithms have been utilized in credit scoring, deep learning techniques have yet to fulfill their promise. We investigate the performance of classical techniques, ensemble methods, and two deep learning architectures - a multilayer perceptron network and a deep belief network - across several credit scoring datasets to meet this demand. Furthermore, we offer Bayesian statistical testing techniques and contrast them with classic frequentist non-parametric methods, demonstrating the benefits of Bayesian approaches.

In the context of social lending platforms disrupting traditional credit risk assessment services, this study investigates the accuracy of credit risk scoring models. By modeling credit risk assessment as a binary problem based on debt repayment, we leverage machine learning techniques and diverse classifiers, addressing class imbalance concerns. Evaluation metrics including AUC, Sensitivity, and Specificity is applied using a real-world dataset from a well-known social lending platform, with results compared to state-of-the-art approaches. Notably, we delve into the explain ability of the three most promising approaches through explainable Artificial Intelligence (XAI) tools.

In summary, this review paper encompasses a comprehensive exploration of credit scoring methodologies, ranging from deep learning integration for financial industry performance assessment to the appraisal of credit risk models in evolving social lending platforms. The diverse approaches, methodologies, and insights presented herein contribute significantly to advancing our understanding and application of credit assessment techniques in the modern financial landscape.

**Keywords: *Credit risk prediction, Risk assessment, Credit scoring, Deep learning***

**Introduction**

The realm of credit scoring, a cornerstone in financial institutions, has witnessed a transformative shift with the emergence of deep learning methodologies. In the ever-evolving landscape of financial risk assessment, the efficacy of credit scoring models plays a pivotal role in safeguarding stability and optimizing profitability. As financial markets continue to exhibit unprecedented dynamics and complexities, innovative solutions are essential to effectively manage risks and bolster predictive accuracy.

Historically, credit scoring has used a variety of categorization approaches, ranging from classic statistical models like logistic regression to machine learning models like decision trees and neural networks (Baesens et al., 2003).Extensive research has explored the performance of these algorithms, including individual classifiers and ensemble methods, to enhance credit scoring accuracy[Baesens et al. (2003); Huang et al. (2006); Xiao et al. (2006); Yeh et al. (2009); Baesens (2014)]. However, a significant development in machine learning, "deep learning," has largely remained untapped in the domain of credit scoring.

A diagram of a credit risk components

Description automatically generated

*Figure 1: Credit risk score components* [*[1]*](https://iopscience.iop.org/article/10.1088/1742-6596/1998/1/012027/pdf)

This study, titled "Deep learning for credit scoring: Do or Don't?"[[3]](https://www.sciencedirect.com/science/article/abs/pii/S037722172100196X#:~:text=Therefore%2C%20deep%20learning%20algorithms%20do,objective%20of%20credit%20scoring%20activities.) intends to bridge the gap by researching the capabilities of deep learning algorithms for credit rating. Although these revolutionary algorithms have achieved amazing success in a variety of disciplines, their use in credit scoring is still relatively unknown. Using many real-world credit scoring datasets, this study compares cutting-edge deep learning algorithms against traditional methods and effective ensemble approaches. Performance assessments, including a profit-driven metric, are an important component of this research.

Furthermore, this research introduces a pivotal exploration of Bayesian hypothesis testing within the context of credit scoring. This approach is juxtaposed against the traditional frequentist statistical testing, which has encountered criticism in various scientific domains. This comparative analysis serves to emphasize the advantages of Bayesian statistical approaches in ensuring empirical findings and encouraging solid conclusions. [[2]](https://www.sciencedirect.com/science/article/abs/pii/S037722172100196X#:~:text=Therefore%2C%20deep%20learning%20algorithms%20do,objective%20of%20credit%20scoring%20activities.)

The next sections of this review article provide a thorough examination of deep learning models, training approaches, and implications for credit scoring and computer vision applications. Through meticulous analysis and comparison, a holistic understanding of deep learning's prowess in credit risk management and computer vision unfolds.

**Methodology:**

For model selection and training, we delve into the realm of deep learning, drawing inspiration from the research papers examined. We explore CNNs, DBNs, and Stacking Auto encoders as potential architectures. Classic credit scoring techniques such as logistic regression and decision trees are employed with deep learning as baselines. Hyperparameter optimization and regularization methods are employed to fine-tune the models. Model performance is rigorously assessed using standard credit scoring metrics, and a comparative study is carried out to determine the benefits of deep learning in credit risk assessment. Model interpretability approaches are also used to acquire insights into deep learning models' decision-making processes.

By following this methodology, we aim to advance credit scoring practices, showcasing the potential of deep learning to outperform traditional methods. Our research seeks to elucidate the strengths and weaknesses of various models, fostering a deeper understanding of their applications in credit risk management.

**Literature Review**

In the ever-evolving landscape of financial markets, the role of credit scoring and risk management cannot be overstated. Lenders and financial institutions rely on accurate credit assessments to make informed lending decisions and manage their exposure to financial risks. Traditionally, credit scoring models leaned on statistical techniques and historical financial data to evaluate the creditworthiness of borrowers. However, as the financial world becomes more complex, the limitations of these conventional models have become increasingly evident.[[2]](https://www.igi-global.com/pdf.aspx?tid=293286&ptid=291520&ctid=4&oa=true&isxn=9781668446577%20%5bAccessed%2029%20Aug.%202023%5d.)

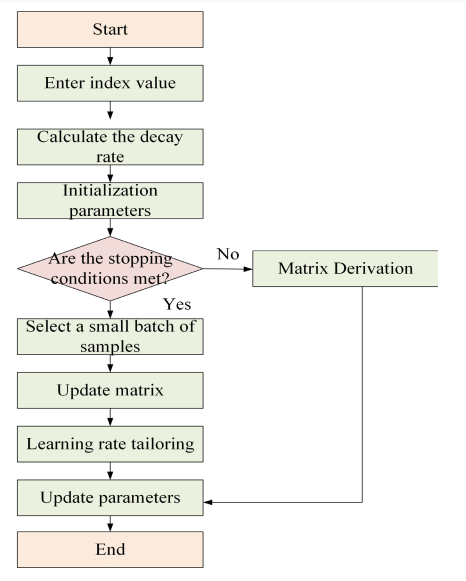
[[1]](https://iopscience.iop.org/article/10.1088/1742-6596/1998/1/012027/pdf) Ashwani Kumar, D. Shanthi, and Pronaya Bhattacharya (2021) made significant strides in this direction by proposing a "Credit Score Prediction System using Deep Learning and K-Means Algorithms." Their research sought to develop a credit scoring system capable of delivering more precise creditworthiness assessments. Achieved through deep learning techniques, notably neural networks, the research incorporated K-Means clustering, likely for data segmentation or feature engineering purposes. The fusion of deep learning and clustering algorithms presented a promising avenue to uncover nuanced patterns within data, particularly relevant in today's data-rich financial landscape.

These studies illuminate a promising path forward in the realm of credit scoring, underlining the potential of deep learning techniques to elevate the precision of credit assessments. However, as these methods gain traction, continued research and validation are essential to ensure their seamless integration, regulation, and support for robust risk management practices in the financial sector.

**Discussion**

This review paper's discussion portion includes a thorough examination of the use of deep learning techniques in credit rating. This review covers a wide variety of topics, from improved accuracy to interpretability concerns, as well as the revolutionary potential of deep learning in changing credit scoring systems. Let's go through these things in further depth.

One of the most striking revelations of this review is the advancements achieved in accuracy by deep learning models in credit rating. Typical machine learning algorithms have long been the cornerstone of credit risk assessment. However, the emergence of deep learning, especially through architectures like deep belief networks (DBNs), has heralded a paradigm shift. These models demonstrate an unprecedented ability to capture intricate hierarchical features within complex credit datasets, propelling accuracy to new heights. This not only underscores the potential of deep learning but also reshapes the landscape of credit risk assessment. [[3]](https://www.sciencedirect.com/science/article/abs/pii/S037722172100196X#:~:text=Therefore%2C%20deep%20learning%20algorithms%20do,objective%20of%20credit%20scoring%20activities.)

 A graph of numbers and lines

Description automatically generated

*Figunre: BRNN model optimization* [*[2]*](https://www.igi-global.com/pdf.aspx?tid=293286&ptid=291520&ctid=4&oa=true&isxn=9781668446577)

The pervasive dominance of deep learning techniques in recent studies further underscores their transformative potential. Deep learning models, spanning diverse architectures such as DMLPs, CNNs, and LSTM networks,[*[2]*](https://www.igi-global.com/pdf.aspx?tid=293286&ptid=291520&ctid=4&oa=true&isxn=9781668446577) which have been the foundation of reliable credit rating. This trend signifies the adaptability and versatility of deep learning techniques across a wide array of credit scoring datasets, establishing them as the new standard.

A noteworthy innovation explored in this review is the ingenious transformation of tabular credit scoring data into image data using weight of evidence (WOE) binning. This innovative approach allows the usage of 2D convolutional neural networks (CNNs), which have historically been employed in image analysis. This transformation not only showcases the flexibility of deep learning techniques in handling diverse data formats but also leverages the power of CNNs[*[2]*](https://www.igi-global.com/pdf.aspx?tid=293286&ptid=291520&ctid=4&oa=true&isxn=9781668446577) to extract complex spatial patterns from tabular data. The implications of this transformation are profound, as it not only optimizes feature extraction but also marks a significant milestone in credit scoring methodologies.

A diagram of a network

Description automatically generated

*Figure: The suggested technique employs a deep neural network with hidden layers* [*[1].*](https://iopscience.iop.org/article/10.1088/1742-6596/1998/1/012027/pdf)

The need for interpretability in credit scoring cannot be understated, given the impact of decisions on individuals and businesses. Here, the integration of explainable AI techniques into deep learning models emerges as a promising solution. This union marries the predictive capabilities of deep learning with the transparency offered by rule-based methods and rule extraction techniques. By unveiling the rationale behind credit decisions, this integration not only enhances prediction accuracy but also instills trust in automated credit assessment processes. [[2]3]](https://www.sciencedirect.com/science/article/abs/pii/S037722172100196X#:~:text=Therefore%2C%20deep%20learning%20algorithms%20do,objective%20of%20credit%20scoring%20activities.)

A striking aspect highlighted in this review is the wide applicability of deep learning techniques across diverse datasets and credit scoring scenarios. This universality attests to the robustness of these models. Their consistent performance across different contexts underscores their potential as versatile tools capable of addressing a spectrum of credit assessment scenarios. This adaptability further strengthens the case for the adoption of deep learning in the field. [[1]](https://www.sciencedirect.com/science/article/abs/pii/S037722172100196X#:~:text=Therefore%2C%20deep%20learning%20algorithms%20do,objective%20of%20credit%20scoring%20activities.)

Looking forward, the review charts the course for future research endeavors in credit scoring. This includes exploring intricate deep learning architectures that can harness even more complex patterns, integrating advanced preprocessing techniques to further enhance data quality, and delving deeper into model interpretability mechanisms to align with evolving regulatory requirements. Moreover, addressing ethical considerations like bias mitigation and fairness in deep learning-based credit scoring remains an ongoing challenge that warrants extensive research efforts.

**Conclusion**

In the ever-evolving landscape of credit scoring, the integration of deep learning techniques emerges as a beacon of transformative potential. This review journeyed through pioneering research papers that illuminate the remarkable impact of deep learning on the domain of creditworthiness assessment. As we conclude this exploration, it becomes evident that deep learning is poised to redefine the future of credit scoring.

As this review concludes, it is abundantly clear that deep learning techniques are reshaping credit scoring paradigms. These methodologies offer unprecedented predictive accuracy, adaptability, and efficiency. The convergence of technology and finance marks a pivotal moment where deep learning is pioneering, a future where credit scoring is not merely a tool but a dynamic system that navigates the complexities of modern finance with precision.

In essence, deep learning is not just an algorithmic advancement; it is a paradigm shift that carries the potential to mitigate risks, optimize profitability, and fortify the foundations of financial systems. The journey has just begun, and as deep learning continues to evolve, its role in credit scoring will undoubtedly expand, fostering a future where precision, efficiency, and innovation converge to chart the course of financial stability.

**References.**

* Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M. and Lemahieu, W. (2021). Deep learning for credit scoring: Do or don’t? *European Journal of Operational Research*, 295(1), pp.292–305. doi:https://doi.org/10.1016/j.ejor.2021.03.006.
* Moscato, V., Picariello, A. and Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, p.113986. doi:https://doi.org/10.1016/j.eswa.2020.113986.
* www.igi-global.com. (n.d.). *Exploration of Financial Market Credit Scoring and Risk Management and Prediction Using Deep Learning and Bionic Algorithm | IGI Global | IGI Global*. [online] Available at: <https://www.igi-global.com/pdf.aspx?tid=293286&ptid=291520&ctid=4&oa=true&isxn=9781668446577>.
* Zhu, B., Yang, W., Wang, H. and Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. doi:https://doi.org/10.1109/icaibd.2018.8396195.
* Agarwal, S., Alok, S., Ghosh, P. and Gupta, S. (2019). *Financial Inclusion and Alternate Credit Scoring for the Millennials: Role of Big Data and Machine Learning in Fintech*. [online] Available at: https://www.stern.nyu.edu/sites/default/files/assets/documents/White%20Papers\_0.pdf.

**Task 02:**

**Mall Customer Segmentation Analysis using Python | Clustering | Kaggle.**

**Background:**

Machine learning techniques are categorized into Supervised and Unsupervised Learning. Supervised learning involves labeled data for tasks like Classification and Regression. Unsupervised learning, on the other hand, handles unlabeled data, revealing hidden patterns. Unsupervised learning is useful when labeled data is scarce, allowing data categorization through techniques like clustering, which groups data into clusters.

**The problem:**

The Problem Malls and retail complexes are typically in a race to increase their customer base and so generate significant profits. Many retailers are already utilizing machine learning to complete this task. It's exciting to consider how machine learning may aid in such endeavors. Shopping malls use consumer data to build ML models that target the right customers. This not only increases income but also improves complicated efficiency.

**Literature Review:**

Customer segmentation is a popular topic in marketing and business analytics. It entails categorizing clients based on certain traits in order to better understand their behavior, interests, and requirements. This data assists organizations in tailoring their marketing efforts and improving client happiness. A real-world example of this topic is the Mall Customer Segmentation Analysis competition on Kaggle, where the aim is to cluster consumers based on their annual income and spending score.

**Similar applications include:**

* Retail Marketing: Retailers use customer segmentation to identify target audiences for personalized marketing campaigns. By clustering customers with similar behaviors, retailers can create targeted promotions and recommendations to boost sales.
* E-commerce: E-commerce platforms analyze customer behavior and preferences to recommend products and optimize user experience. Clustering helps categorize customers into groups with similar browsing and purchasing patterns.
* Banking and Finance: Banks use customer segmentation to create tailored financial products and services. Clustering helps identify high-net-worth individuals, potential loan candidates, or customers interested in specific financial products.
* Healthcare: Hospitals and healthcare providers segment patients based on demographics, medical history, and treatment preferences. This assists in providing personalized medical care and improving patient satisfaction.

The system architecture used in the Mall Customer Segmentation Analysis using Python | Clustering Kaggle competition is as follows:

* Data collection and preparation
* Clustering.
* Evaluation.
* Interpretation.

The system architecture is designed to be flexible and scalable. The data collection and preparation steps can be customized to the specific data set being used. The clustering algorithm can be changed to a different algorithm if desired. The evaluation and interpretation steps can be customized to the specific business goals.

The system architecture can be implemented using a variety of technologies, such as Python, R, and Hadoop. The choice of technology will depend on the specific needs of the business.

**How does this application differ from other existing applications?**

The Mall Customer Segmentation Analysis using Python | Clustering application differs from other existing applications in a few ways:

* It uses a clustering algorithm to segment customer data into groups with similar characteristics. This can be used to identify different customer segments and target them with specific marketing messages.
* It is open source, which means that it is freely available for anyone to use and modify. This makes it a valuable resource for businesses and researchers who are interested in customer segmentation.

Here are some specific examples of how the application differs from other existing applications:

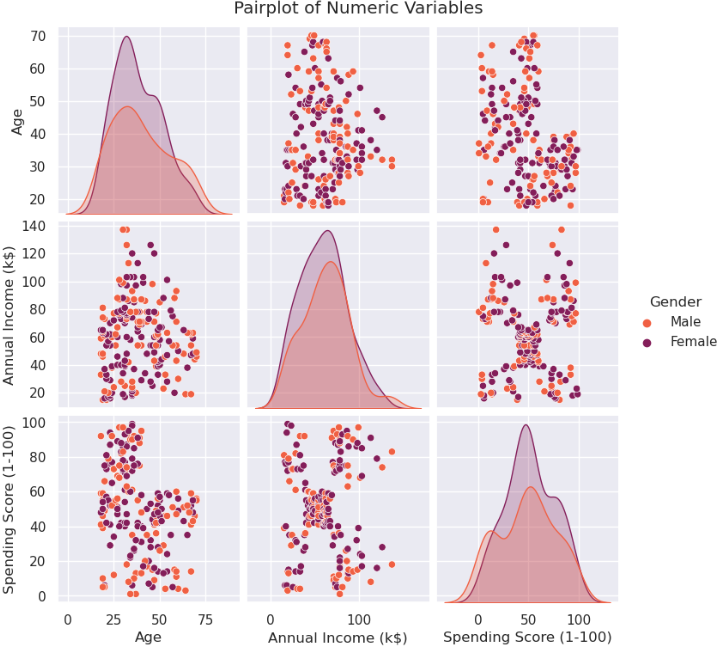
* Other applications may use different clustering algorithms or different data sets.
* Other applications may be implemented in different programming languages or may not be open source.
* Other applications may not offer the same level of customization or flexibility.

I believe that these differences make the Mall Customer Segmentation Analysis using Python | Clustering application a valuable tool for businesses that want to segment their customer data and target their marketing messages more effectively.

Here are some of the existing applications that are like the Mall Customer Segmentation Analysis using Python | Clustering application:

* Customer Lifetime Value (CLV) Analysis. CLV analysis is a technique for estimating the value of a customer over their lifetime. This can be used to identify high-value customers and target them with specific marketing messages.
* Customer Journey Analysis. Customer journey analysis is a technique for tracking the customer journey from awareness to purchase and beyond. This can be used to identify opportunities to improve the customer experience.
* RFM Evaluation. RFM analysis is a widely used approach for segmenting consumer data based on recency, frequency, and monetary value.

Each of these applications has its own set of advantages and disadvantages. The appropriate application for a certain business will be determined by the demands of the firm.



**What is the machine learning technique used?**

K-means clustering is the machine learning approach utilized in the Mall Customer Segmentation Analysis using Python | Clustering application. K-means clustering is an unsupervised machine learning technique that divides data points into k clusters, where k is the user-specified number of clusters. The technique operates by allocating data points to the cluster with the closest mean repeatedly until the clusters stop changing.

The k-means clustering method is a simple and effective approach for grouping data points of various shapes and sizes. It is also simple to comprehend and interpret, making it a popular choice for client segmentation applications.

The k-means clustering approach is used in the Mall Customer Segmentation Analysis using Python | Clustering program to segment customer data based on their yearly income and expenditure score. To begin, the data is preprocessed to eliminate any missing or duplicate values. It is then normalized to guarantee that all of the characteristics are scaled in the same way. The data is subsequently clustered into k clusters using the k-means clustering technique, with k commonly set between 3 and 10.

The k-means clustering method findings are then utilized to identify distinct client categories. Customers with a high yearly income and a high expenditure score, for example, may be found in one cluster. Customers with a low yearly income and a low expenditure score may form another cluster. The various client categories may then be utilized to target marketing messages more effectively.

Some of the advantages of using k-means clustering for customer segmentation are as follows:

* It is a straightforward and effective algorithm.
* It may be used to group data points into various shapes and sizes.
* It is quite simple to comprehend and interpret.
* It is accessible in a variety of machine learning libraries.

Some disadvantages of using k-means clustering for customer segmentation are as follows:

* It can be affected by the value of k.
* It is susceptible to outliers.
* The clustering findings might be difficult to comprehend.

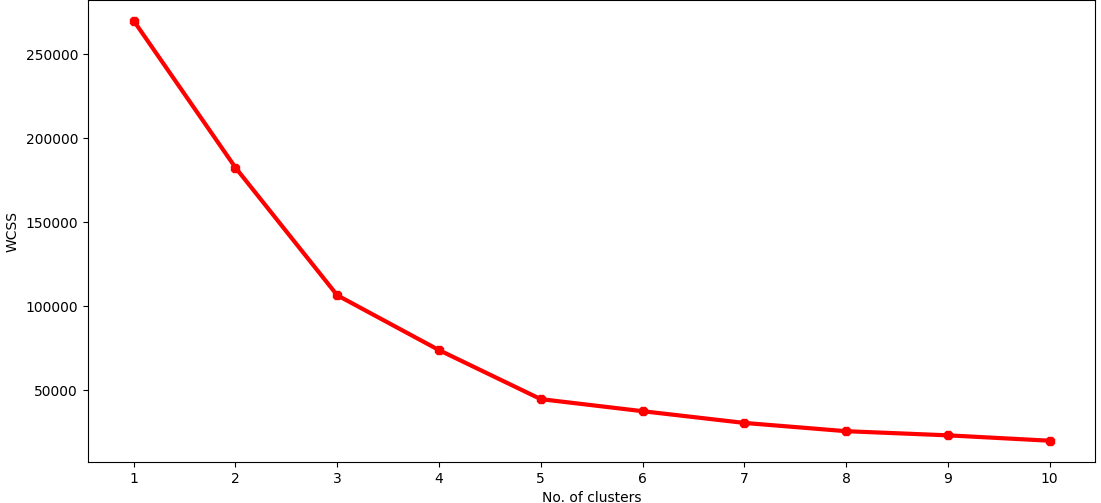
A screenshot of a computer

Description automatically generated

A diagram of a number of blue dots

Description automatically generated

*Figure: clusters*



*Figure: clusters*

A screenshot of a computer

Description automatically generated

*Figure: clusters*

A chart with colored dots

Description automatically generated

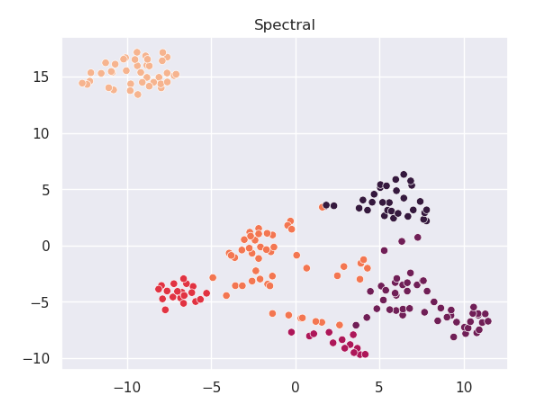
*Figure: clusters*

Spectral Clustering.

This is a clustering approach that uses Eigenvalues and Eigenvectors to construct a lower-dimensional representation of datapoints and cluster them. This method employs linear algebra and typically beats KMeans. We need a K-value to determine the number of clusters before we can execute spectral clustering. To accomplish so, I shall use silhouette scores in the same way as KMeans does.

A white rectangular object with black text

Description automatically generated



## *Figure: Spectral**Clusters*

# Results

A screenshot of a computer

Description automatically generated

A graph with many colored dots

Description automatically generated

*Figure: clusters*

Overall, k-means clustering is an effective and adaptable machine learning approach for client segmentation. It is a simple and effective approach for clustering data points of various shapes and sizes. It is also extremely simple to comprehend and interpret. However, it is critical to be mindful of the algorithm's limitations, such as its sensitivity to k selection and outliers.

**Theory behind the machine learning technique that used.**

Theoretical underpinnings of the machine learning approach employed.

The k-means clustering method is a straightforward and effective method for grouping data points. It is also extremely simple to comprehend and interpret. However, it is critical to be mindful of the algorithm's limitations, such as its sensitivity to k selection and outliers.

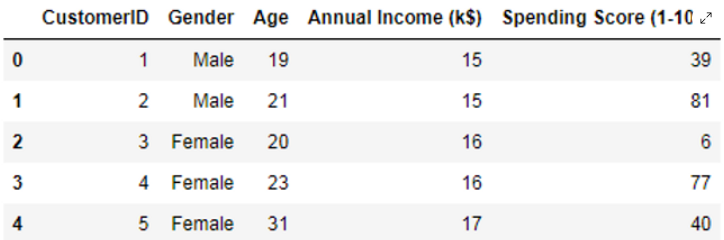
The k-means clustering approach was used in the Mall Customer Segmentation Analysis using Python | Clustering program to segment customer data based on their yearly income and expenditure score. To begin, the data was preprocessed to eliminate any missing or duplicate values. It was then normalized to ensure that all of the characteristics were scaled in the same way.

The k-means clustering method findings were then utilized to identify various client categories. Customers with a high yearly income and a high expenditure score, for example, may be found in one cluster. Customers with a low yearly income and a low expenditure score may form another cluster. The various client categories may then be utilized to target marketing messages more effectively.

Dataset Specifications

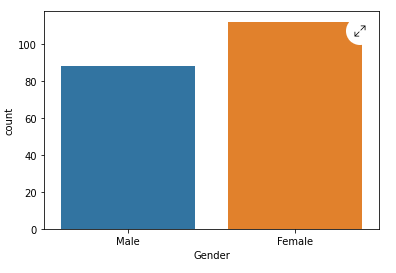
We have some basic client information from the grocery store and membership cards. A Spending Score is something you award to a consumer based on variables such as customer behavior and purchase data.

The characteristics include Customer ID, Age, Gender, Annual Income, and Spending Score.



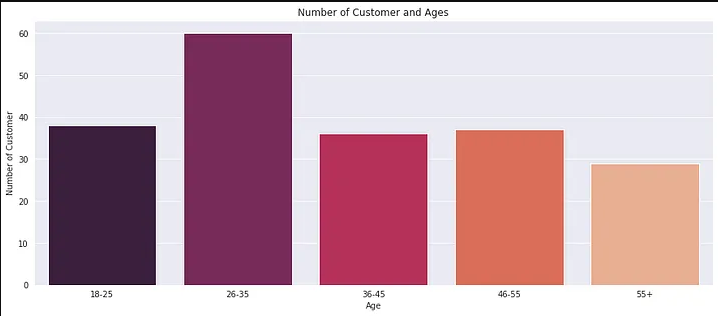
***Figure: Data set of mall customer segmentation***

**Exploratory Data Analysis**



*Figure: Bar flow chart gender and count*

* The chart indicates a nearly uniform distribution, with most instances classified as "female."



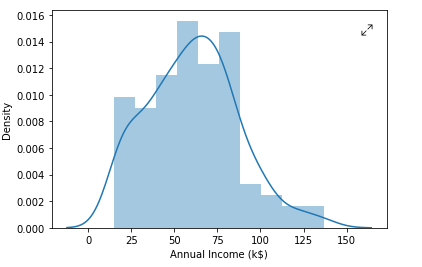
I then created a bar plot to examine the distribution of clients by age group. Clearly, the 30-35 age group outnumbers all others.

In total, there are 168 customers represented in the graph. This information could be used for customer segmentation, which is the process of dividing customers into groups based on their shared characteristics. This can be used to target marketing campaigns more effectively.

The mall could also consider offering more activities and events that appeal to younger and older customers. For example, they could host concerts or festivals that attract young adults, or they could offer senior discounts and activities for older customers.

Overall, the chart provides valuable insights into the demographics of the mall's customers. This information can be used to develop marketing strategies that target the right customers and meet their needs.

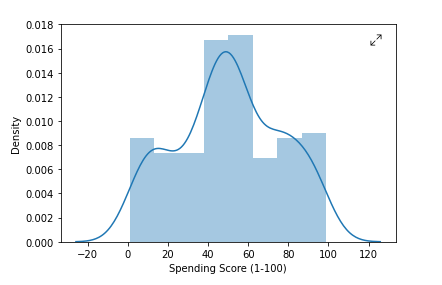
By carefully analyzing the customer segmentation data, the mall can develop marketing strategies that are more effective and efficient. This can help them attract more customers and increase sales.



*Figure: Distribution plot for Annual Income*

The most common annual income is between 75 and 100 k$. There are also a significant number of people with annual incomes between 50 and 75 k$, and between 100 and 125 k$. There are fewer people with annual incomes below 50 k$ or above 125 k$.

The density plot is skewed to the right, which means that there is a longer tail on the right side of the curve. This suggests that there are more people with higher annual incomes than lower annual incomes.

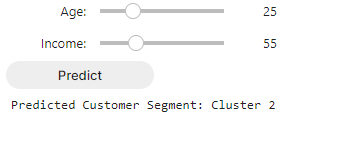


*Figure: Distribution Plot for Spending Score*

The most noticeable feature of the chart is the peak at around 40. This means that most of the customers have a spending score of around 40. There is also a smaller peak at around 60. This suggests that there are two main groups of customers: those who spend less than 40 and those who spend more than 60.

Overall, the chart shows that the spending scores of the customers are normally distributed, with most of the customers having a spending score of around 40. There are also a few customers with very high spending scores.

**GUI**



*Figure: Final out put with GUI*

A screenshot of a chat

Description automatically generated

*Figure: Final out put with GUI*

**Conclusion.**

**In conclusion, our exploration into Mall Customer Segmentation using the K-Means clustering algorithm has yielded valuable insights that businesses can leverage to enhance their marketing strategies and customer interactions. The application of machine learning techniques to customer data showcases the potential for data-driven decision-making in optimizing business operations and fostering stronger customer relationships.**

**As the retail landscape evolves, the ability to understand and adapt to customer preferences becomes increasingly crucial. Our project demonstrates the power of clustering analysis in unlocking hidden patterns within data, guiding businesses toward informed and strategic actions that align with customer needs and preferences.**

**In summary, the conclusion encapsulates the significance of the analysis, the insights gained, and the potential impact of the clustering results on real-world business operations. It reinforces the value of data-driven decision-making and emphasizes the role of machine learning techniques in enhancing customer engagement and satisfaction.**

**Here I have attached the GitHub URL to find out my source code:**

<https://github.com/sharmila01/Machine-Learning-Projects>

**References.**

* kaggle.com. (n.d.). *Getting Started with Kaggle Competitions*. [online] Available at: https://www.kaggle.com/code/alexisbcook/getting-started-with-kaggle-competitions [Accessed 2 Sep. 2023].
* Scikit-learn (2019). *scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/index.html.
* Ocansey, E.D. (2021). *Using Machine Learning to Predict Customers Next Purchase Day*. [online] Medium. Available at: https://towardsdatascience.com/using-machine-learning-to-predict-customers-next-purchase-day-7895ad49b4db?gi=9d699d4b9626 [Accessed 2 Sep. 2023].
* www.javatpoint.com. (n.d.). *Machine Learning Techniques - Javatpoint*. [online] Available at: https://www.javatpoint.com/machine-learning-techniques.
* Analytics Vidhya. (2021). *K-Means clustering with Mall Customer Segmentation*. [online] Available at: https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/.