PAPER NAME **1st assignment-5.docx**	AUTHOR **-**
WORD COUNT **955 Words**	CHARACTER COUNT **5590 Characters**
PAGE COUNT **7 Pages**	FILE SIZE **158.8KB**
SUBMISSION DATE **Jul 20, 2023 9:20 PM GMT+5:45**	REPORT DATE **Jul 20, 2023 9:21 PM GMT+5:45**

● 19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- Crossref database
- 19% Submitted Works database
- 0% Publications database
- Crossref Posted Content database

Full Name: Sushant Pandit

Student ID: 2065998

Group: L6CG2

Lecturer: Simon Giri

Tutor: Sunita Parajuli

1 Long Question:

Q. As a data scientist at e-commerce, list potential use cases where machine learning (Classification and Regression) can be applied in the company. Identify at least 2 actual areas (one each for classification and Regression) where this technique can be used and explain which algorithms can be employed to solve these problems.

Machine learning has been a founding innovation in e-commerce since its inception, and its benefits have been widely recognized. The application of machine learning techniques has improved nearly every element of e-commerce. (Haponik, 2021)

Working as a data scientist for an e-commerce firm gives you many chances to employ machine learning techniques such as classification and regression. There are several possible use scenarios in which these technologies can be efficiently used. Here is actual two areas where these techniques can be effectively utilized:

Classification: Fraud Detection

With the development in online transactions enabled by various payment methods such as credit/debit cards, online banking, E-Sewa, Khalti, and so on, there has also been an increase in fraudulent activity. Detecting and combating fraud has become a critical duty for e-commerce enterprises to protect their consumers and company interests. Payment fraud has become a widespread problem, with counterfeiters stealing cards and creating counterfeit copies to exploit for personal benefit. (Intellipaat, 2023)

Machine learning classification algorithms, which examine various data aspects and trends, play a critical role in spotting possibly fraudulent transactions. They excel in precisely classifying transactions as legal or suspect, assisting in the identification and prevention of fraudulent activity in the e-commerce ecosystem. Algorithms that can be used for fraud detection are:

a. Supervised Learning for Fraud detection:

The model undergoes training using labeled data, with each transaction classified as 'fraud' or 'non-fraud.' A large amount of labeled data is utilized to train the supervised learning model, allowing it to produce correct results. The models, once trained, can predict the class labels of fresh and previously unknown transactions.(Intellipaat, 2023)

b. Unsupervised Learning for Fraud detection:
These models are intended to find uncommon abnormalities or patterns that may not have been noticed previously. They use a self-learning system to identify hidden patterns in transactions. The models strive to uncover parallels and dissimilarities in the occurrence of transactions by evaluating existing data, boosting their capacity to detect possible fraudulent acts. (Intellipaat, 2023)

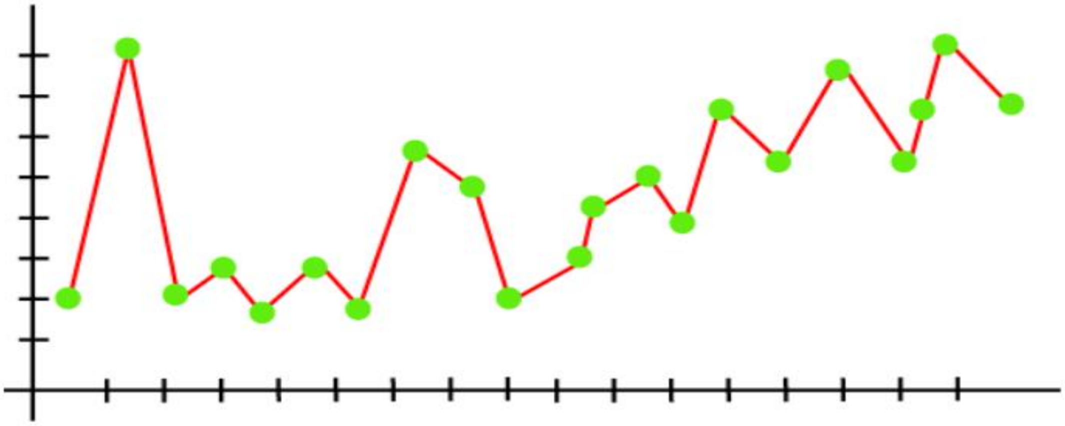
2 Short Question:

2.1 Overfitting:

Q. What are overfitting and underfitting? Why is it a problem in machine learning? Explain the concepts with the help of examples.

Overfitting and underfitting are two common challenges that develop when the accuracy of a model is unsatisfactory in machine learning. The basic goal of every machine learning model is to attain good generalization, which means that it should be capable of producing correct results by adapting to previously unknown input data.

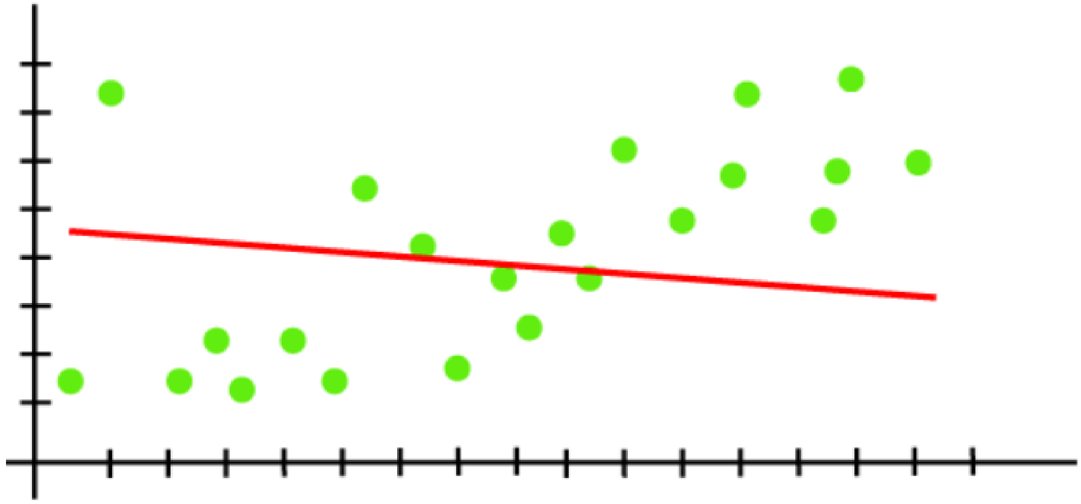
Overfitting occurs when a machine learning model becomes highly complicated or overly customized to the training data, collecting noise or irrelevant patterns and reducing the model's efficiency and accuracy. The model's performance is defined by low bias and large variation in this phenomenon.



The graph shows that the model tries to incorporate all the data points from the scatter plot. While this appears to be a thorough method, it is not always efficient. The basic goal of regression modeling is to determine the best fit line, however in this scenario, there is no obvious best fit line, resulting in higher prediction errors. The ways to reduce overfitting are:

- Cross- Validation
- Removing features
- Early stopping the training

When a machine learning model fails to capture the underlying trend or patterns in the data, this is referred to as underfitting. It misses the crucial connections between the characteristics and the goal variable. Underfit models often have significant bias and low variance, resulting in poor performance on both training and test data.



Based on the provided graph, the model seems to be incapable of capturing the data points presented in the plot effectively. To avoid underfitting, there are two potential approaches:

- Increase the training time of the model to allow it to learn from the data more thoroughly.
- Enhance the model's performance by increasing the number of features, enabling it to better represent the underlying relationships within the data. (JavaTpoint, 2021)

2.2 Training-Machine Learning:

Q. We used Gradient descent to optimize the training of Linear and Logistic Regression. Explain in short your understanding of Gradient Descent.

Gradient descent is a popular machine learning optimization approach that is used to iteratively reduce the cost function by altering parameters in the opposite direction of the negative gradient. Its major goal is to determine the best set of variables by computing the gradient and amplitude of the sharpest climb. This approach is notably useful in Linear and Logistic Regression Models, where the goal is to minimize the provided function to improve model performance.

1. Initialization: The algorithm starts by initializing the model's parameters or coefficients with some values.
2. Calculate Gradient: The gradient of the cost function is calculated with respect to each parameter. The directions and magnitude of the steepest increase in cost function represents gradient.
3. Update parameters: For each parameter, new values is obtained by subtracting a small fraction of the gradient from the current values. This step is repeated iteratively until convergence is reached.
4. Convergence: As the parameter updates iteratively, gradually moving towards the minimum of the cost function. When the change in the cost function becomes small, indicating that the optimal parameters have been found, convergence is achieved.
5. Optimized Model: Once the algorithm converges, the model's parameters are optimized, and the model is ready to make predictions on new unseen data.

Overall, Gradient descent is a powerful and widely used optimization algorithm that plays a fundamental role in training machine learning models by iteratively finding the optimal parameters values that minimize the cost function and improve the model's performance. (Crypto1, 2020)

Works Cited

Crypto1. (2020, October 2). *How Does the Gradient Descent Algorithm Work in Machine Learning?* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/how-does-the-gradient-descent-algorithm-work-in-machine-learning/#:~:text=A.-,GradientHaponik, A. (2021, June 23). *Machine Learning in Ecommerce: 8 Best Use Cases | Addepto*. Retrieved from Addepto: https://addepto.com/blog/best-machine-learning-use-cases-ecommerce/

Intellipaat. (2023, June 28). *Fraud Detection Algorithms | Fraud Detection using Machine Learning*. Retrieved from Intellipaat: https://intellipaat.com/blog/fraud-detection-machine-learning-algorithms/?US#no3

JavaTpoint. (2021). *Overfitting and Underfitting in Machine Learning - Javatpoint*. Retrieved from Javatpoint: https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning

● **19% Overall Similarity**

Top sources found in the following databases:

- 2% Internet database
- Crossref database
- 19% Submitted Works database
- 0% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	**The University of Wolverhampton on 2023-07-20** Submitted works	8%
2	**The University of Wolverhampton on 2023-05-08** Submitted works	3%
3	**Wilmington University on 2020-10-09** Submitted works	2%
4	**Liverpool John Moores University on 2022-11-28** Submitted works	2%
5	**Kennesaw State University on 2023-03-20** Submitted works	1%
6	**The University of Wolverhampton on 2023-05-08** Submitted works	1%
7	**The University of Wolverhampton on 2023-05-08** Submitted works	1%
8	**The University of Wolverhampton on 2023-05-08** Submitted works	<1%