1: introduction to Regression

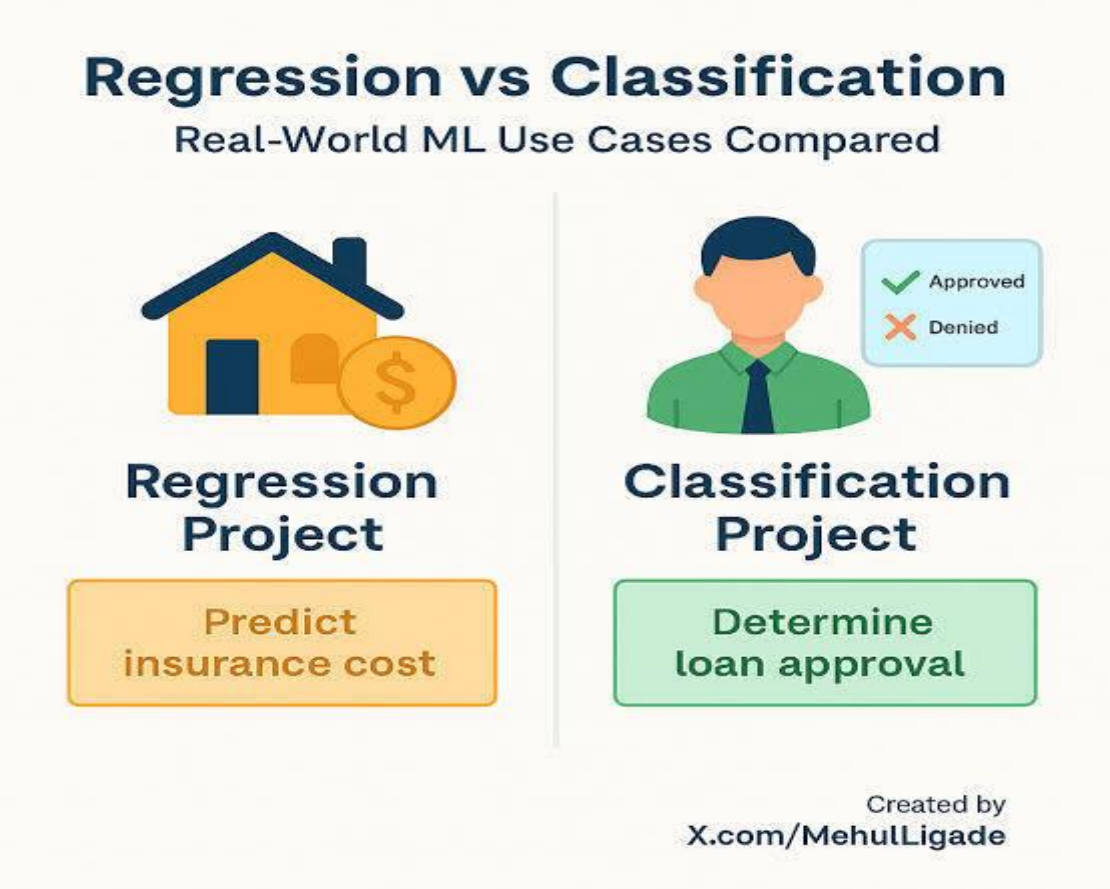
What is regression in Machine Learning?

is a type of supervised learning used to predict continuous, numerical output values (like house prices, age, or stock market trends) based on input features

Regression in machine learning is predict a continuous outcome (y) based on the value of one or more predictor variables (x)

Linear regression is probably the most popular form of regression analysis because of its ease-of-use in predicting and forecasting

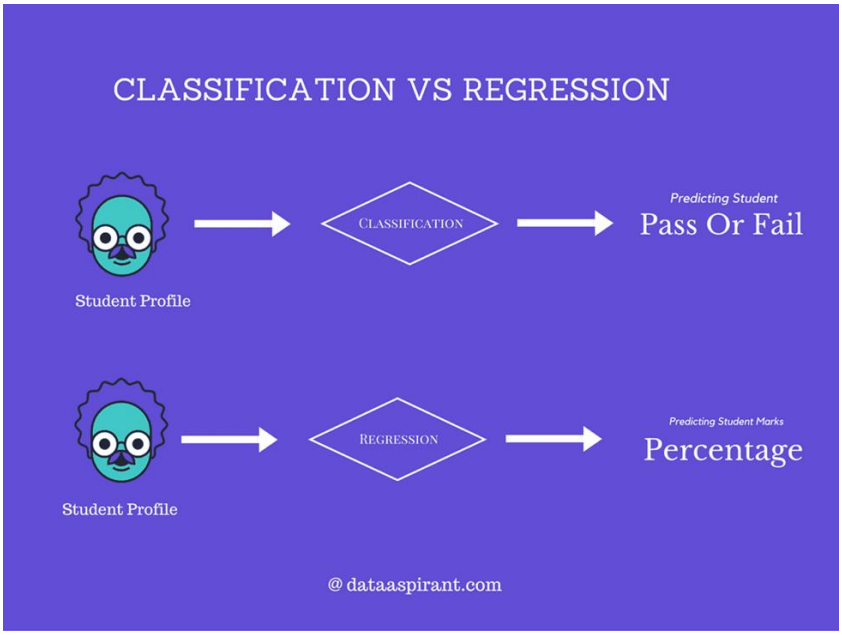
How is Regression different from classification?



The fundamental difference is that classification predicts a discrete label or category spam or not spam, while regression predicts a continuous, real valued number

Example: Regression is used when the output target variable is continuous, the goal is to predict a numeric value

Example: Classification is used when the output target variable is categorical, the goal is to assign a category or class to input data



Difference:

Regression: answers how many numbers

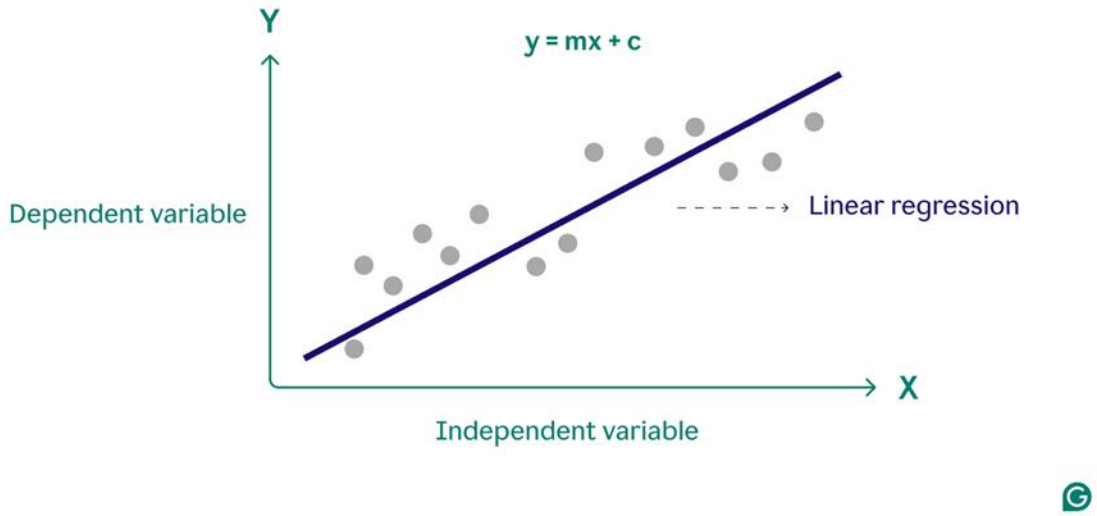
Classification: answers which category are there

Give one real-life example of regression and one of classification?

Type	Real-life Example	Target Type
Regression	car resale value mileage, age, brand, and condition	Continuous value
Classification	Credit card fraud detection – classifying transactions as fraudulent or legitimate	Discrete label
Classification	student performance – classifying students as pass or fail based on study hours, attendance, and past scores	Discrete label
Regression	electricity in a household based on temperature, appliances used, and time of day	Continuous value

2. Types of Regression

1: Linear Regression: Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events, It is a statistical method used in data science and machine learning for predictive analysis



Advantages:

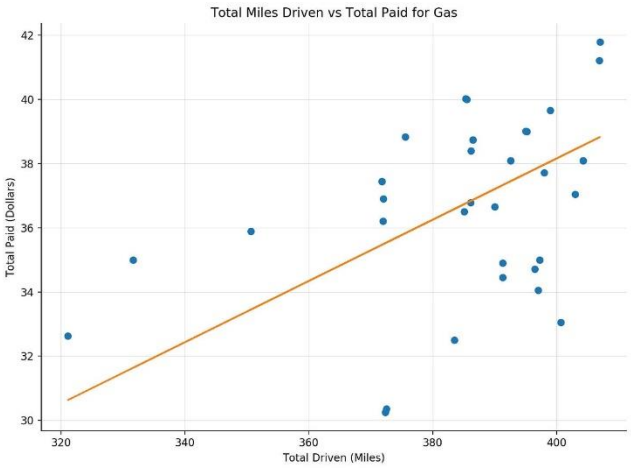
1. Simple implementation: The system does not require advanced setup, Additional maintenance is also minimal
2. User friendly: The system is straightforward, unlike complicated models like neural networks that obfuscate rather than clarify
3. Efficient computation: The system can even sustain big data and is not expensive itself
4. Real time performance: Instant updates are possible, meaning perishable data can also be analyzed without time loss

Limitations:

1. Linearity Assumption: Assumes a linear relationship between input and output, fails if the relationship is nonlinear
2. Sensitivity to Outliers: Outliers can significantly affect the model's predictions

3. Limited Complexity: Cannot capture complex patterns in data like neural networks or tree-based models
4. Multicollinearity Issues: Highly correlated features can distort the interpretation of coefficients

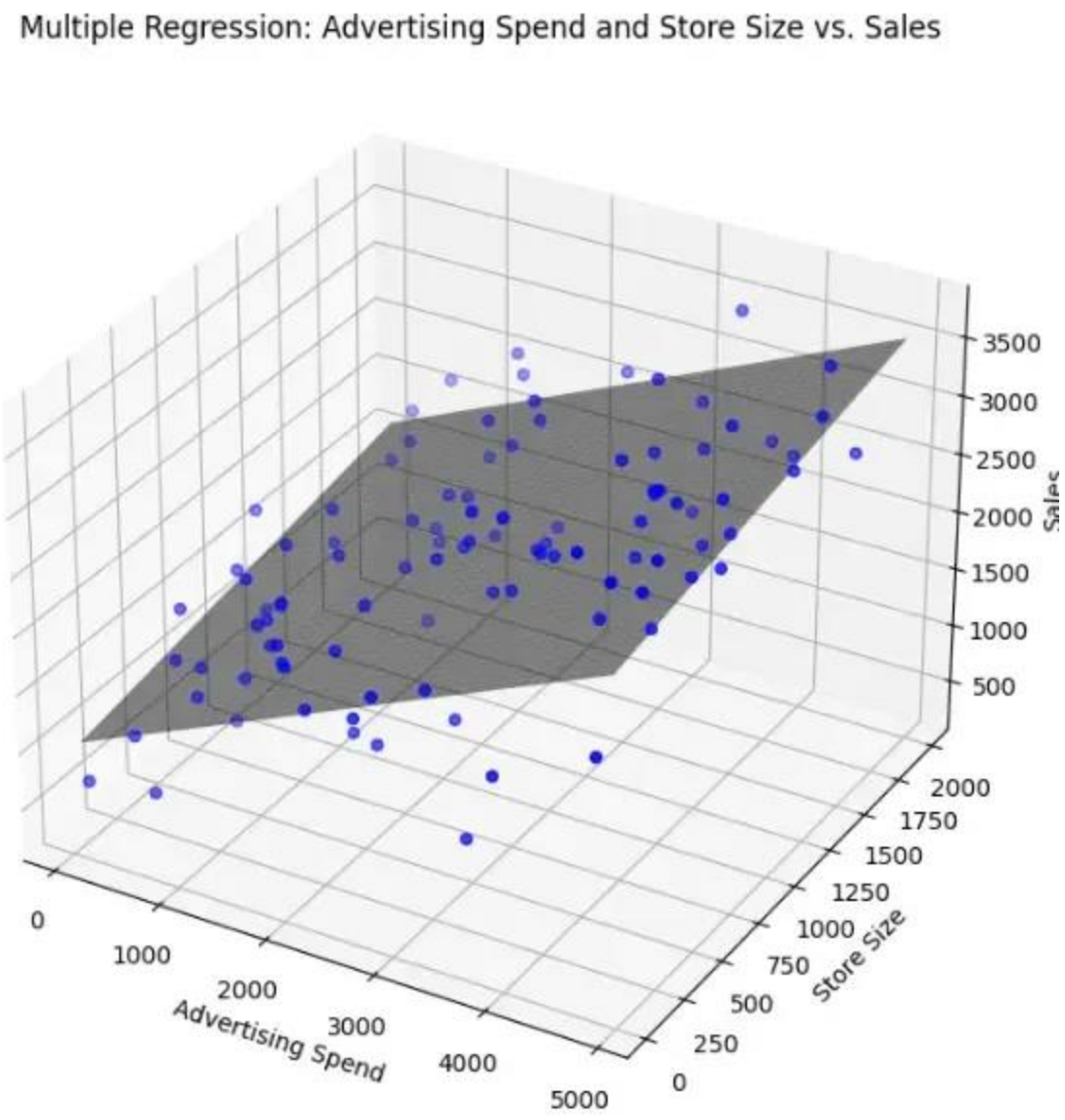
Advantages	Limitations
Simple and easy to implement	Assumes a linear relationship (not good for nonlinear data)
Easy to interpret	Sensitive to outliers
Scalable for large datasets	Cannot capture complex patterns
Suitable for real-time prediction	Multicollinearity can distort coefficients
Good baseline model	Not suitable for categorical outputs



One real world use case:

A company aims to forecast monthly sales revenue based on factors like advertising spend, number of promotions, and seasonal trends

2. Multiple Linear Regression MLR: Models the relationship between two or more independent variables and a continuous dependent variable

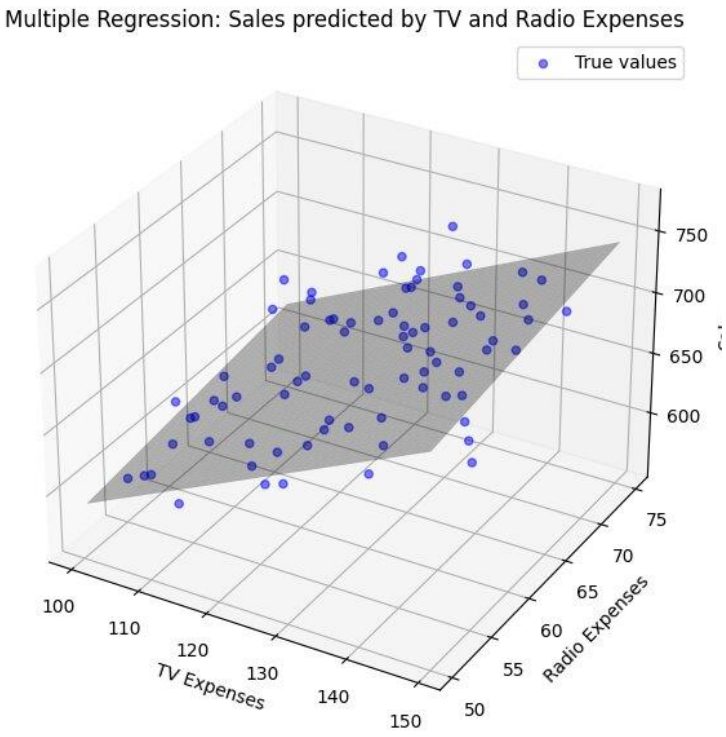


Advantages:

1. Handles multiple independent variables simultaneously
2. Easy to interpret coefficients for each predictor
3. Useful for predicting continuous outcomes

Limitations:

1. Assumes linear relationships between inputs and output
2. Sensitive to outliers
3. Multicollinearity (highly correlated predictors) can distort results



One real world use case:

Predicting sales of Tv & Radio based on height, weight, and exercise

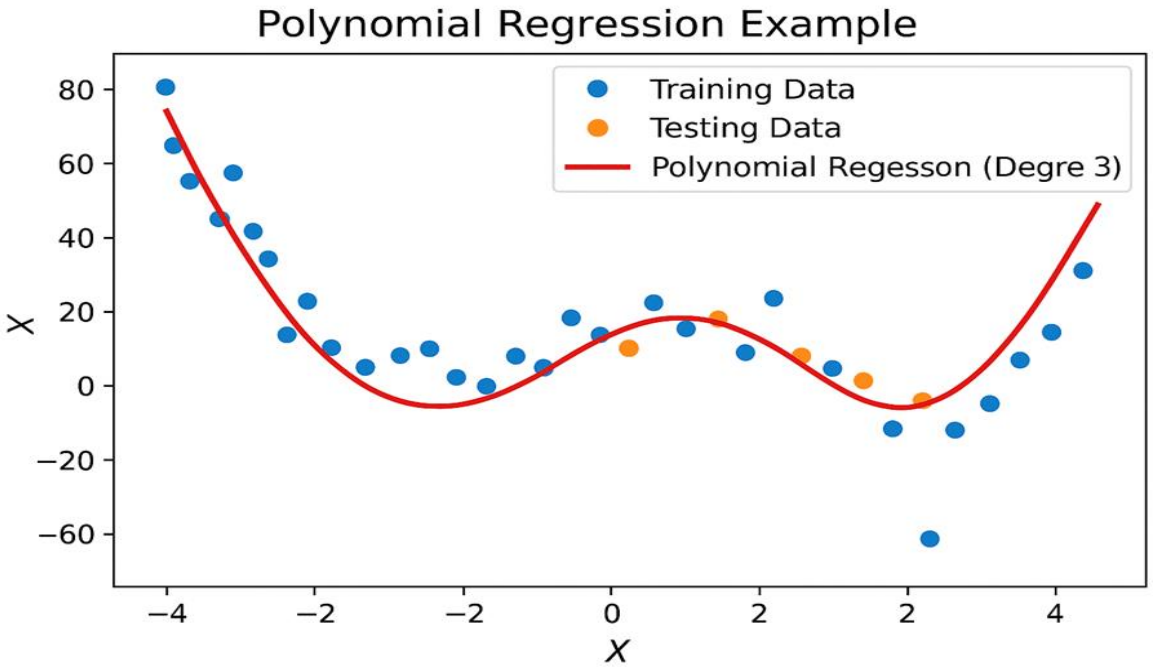
Waxa jira kuwo kale

3. Logistic Regression

4. Ordinal Regression

5. Multinomial Logistic Regression

3. Polynomial regression: is a statistical method that models the relationship between a dependent variable and one or more independent variables as an n-th degree polynomial

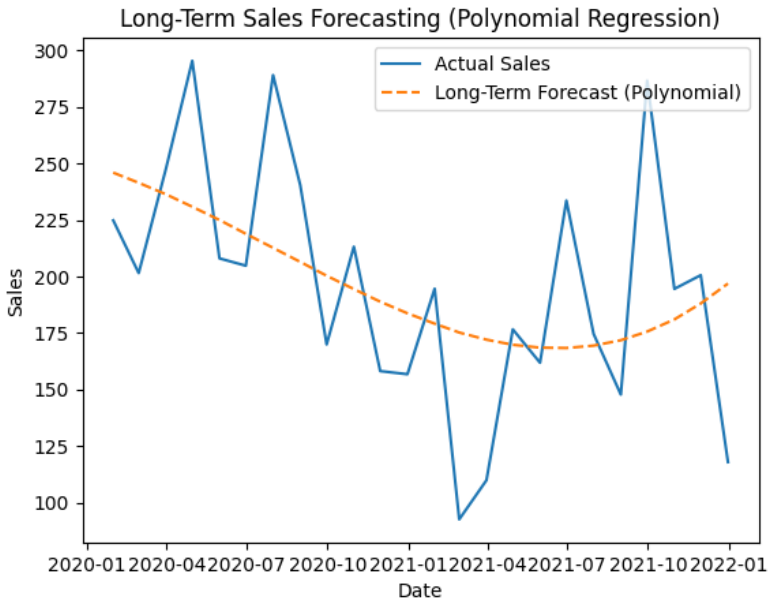


Advantages

1. Capture Non-Linearity: Being able to model curved relationships while linear regression cannot do so
2. Flexibility: Polynomial regression can capture relationships from linear to highly nonlinear which can be crucial when performing real world data set that do not follow simple linear relationships
3. Improved Accuracy: The more complex a model is, the more complex the predictions will be, and thus polynomial regression will outperform simple linear regression
4. Interpretability: Polynomial regression models are fairly simple to explain. The relationships from the coefficients can help the practitioners and researchers understand which variable are the crucial ones driving the data set
5. Flexible Fit: Complex data relationships can be better fit by the model by using a polinomial function of higher degree
6. Still Linear in Parameters: Although the relationships between the variables are nonlinear, the model can be easily estimated through linear regression
7. Predictive Power: Forecasting with non linear relationships, such as growth curves and seasonal patterns, is very effective.

Limitations

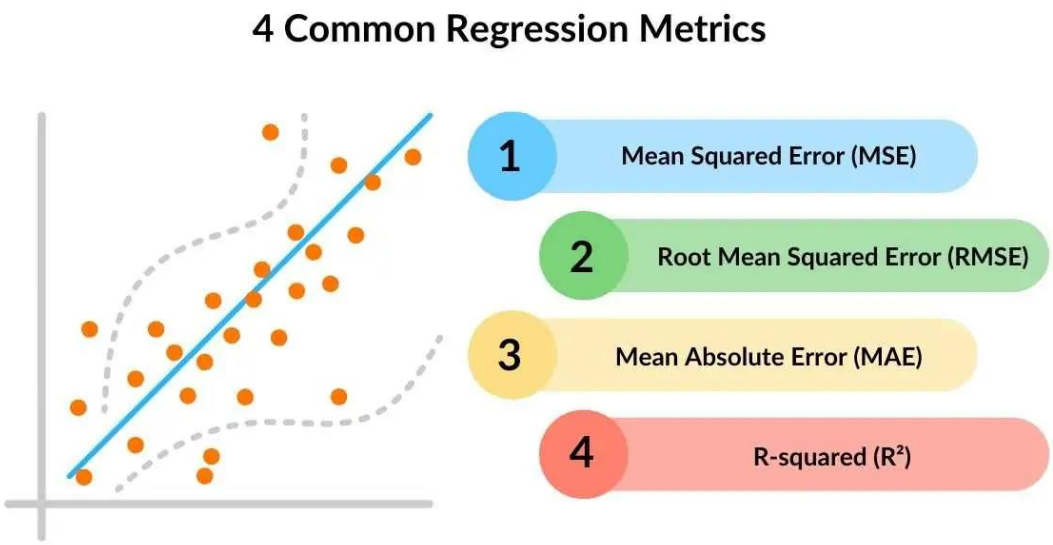
1.Overfitting Risk: Noise could be captured by real trends on high-degree polynomial curves
2.Overfitting: Too close of a fit on training data while losing generalization capability is known as overfitting. Such higher-degree polynomial models would require careful selection and the application of regularization techniques to these splits
3.Computational Complexity: Complexity increases as the degree of the polynomial increases. Polynomials with exceedingly high degrees could be computationally burdensome and may fail to yield any significant increases on model performance
4.Data Requirements: In polynomial regression, high-degree polynomial regression models have a certain amount of data requirements to be considered. Such small datasets fail to estimate the polynomial segments
5.Extrapolation Issues: Inaccurate predictions fall outside the observed range
6.Interpretability: Coefficients in higher degrees are comparatively more difficult to reason than lower degree
7.Sensitive to Outliers: Curve fit can be affected by extreme values



One real world use case:

Researchers aim to understand long sales in polyonominal regression in the sales to date making polynomial regression a suitable modeling approach

3. Regression Metrics



1. MAE_Mean Absolute Error: is measures the average absolute difference between the predicted values and the actual values, on average, how far off your predictions are from the true values, ignoring the direction of the error (no negative/positive)

Example: If MAE = 5, your predictions are off by 5 units on average

2. MSE – Mean Squared Error: is measures the average of the squared differences between predicted and actual values, it penalizes larger errors more than smaller ones because errors are squared

Example: If one prediction is very far from the actual value, it will increase the MSE significantly

3. RMSE – Root Mean Squared Error: RMSE is simply the square root of MSE, like MSE, it penalizes larger errors, but now the units are the same as the original data, making it easier to interpret

Example: If predicting house prices in $1000s and RMSE = 10, your predictions are off by about $10,000 on average

4. R^2 – Coefficient of Determination: R^2 measures how much of the variation in the dependent variable is explained by your model

1 = perfect prediction

0 = model predicts no better than the mean

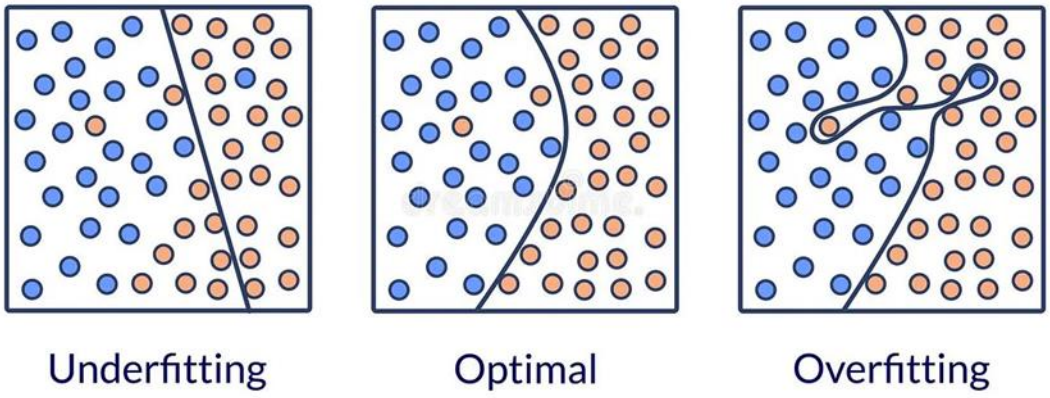
< 0 = model is worse than predicting the mean

Example: If $R^2 = 0.95$ your model explains 95% of the variation in the data

Comparison table showing their differences:

Type	**Relationship Shape**	**Example Features**	**Best For**
Linear Regression	Straight line	Years of Experience Salary	Simple, single-variable predictions
Multiple Linear Regression	Multi-dimensional plane	Age, Education, Experience Salary	Predicting outcomes with multiple factors
Polynomial Regression	Curved line	Time, Time^2 Plant Growth	Non-linear growth or trends
Logistic Regression	S-curve sigmoid	Visits, Clicks Purchase Yes/No	Binary classification tasks
Ordinal Regression	Ordered categories	Survey Responses Disagree to Strongly Agree	Predicting ranked outcomes
Multinomial Logistic Regression	No specific order	Student Program Choices Vocational/Sports/Academic	Predicting outcomes with multiple nominal categories

4. Underfitting and Overfitting



Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data When data scientists use machine learning models for making predictions, they first train the model on a known data set Then, based on this information, the model tries to predict outcomes for new data sets. An overfit model can give inaccurate predictions and cannot perform well for all types of new data

Underfitting occurs when the model is too simple to capture the underlying pattern in the data other words, the model is not complex enough to represent the true relationship between the input and output variables, Underfitting can occur when the model is too simple or when there are too few features relative to the number of training examples consider a simple linear regression problem where we want to predict the height of a person based on their weight if we use a linear model to fit the data, we may not capture the curvature in the relationship between weight and height. In this case, the model is too simple to capture the true relationship between the input and output variables

What causes overfitting, especially in polynomial regression?

It's used polynomial degrees to capture the relationship between features and target, High degree polynomials very wiggly curve that tries to fit all training points exactly, this causes the curve to work well on training data, but poorly on test data (large generalization error)

1. Model Complexity: Highly complex models with numerous parameters can capture noise in the training data, leading to overfitting
2. Data Sparsity: In recommendation systems, user-item interaction matrices are often sparse, making it challenging to generalize patterns
3. Imbalanced Data: When certain user or item categories dominate the dataset, the model may overfit to these categories while neglecting others
4. Inadequate Regularization: Without proper regularization techniques, models are more likely to overfit
5. Overtraining: Training a model for too many epochs can lead to overfitting, as the model starts to memorize the training data
6. Poor User Experience: Overfitted models may provide irrelevant or overly specific recommendations, frustrating users
7. Reduced Engagement: If users find the recommendations unhelpful, they are less likely to engage with the platform
8. Financial Losses: In e-commerce, poor recommendations can lead to lost sales and reduced customer loyalty
9. Bias Amplification: Overfitting can exacerbate existing biases in the data, leading to unfair or discriminatory recommendations
10. Operational Inefficiencies: Overfitted models may require frequent retraining, increasing computational costs and resource usage

Example:

Degree 2 or 3 Usually captures a general trend

Degree 10 Covers all points, and can be an oscillating curve that makes no sense

Give 2–3 methods to prevent overfitting

1. Hold-out: Split data into training and testing sets example 80/20 to check generalization
2. Cross-validation: Split data into k folds, train/test multiple times so all data is used
3. Data augmentation: If more data can't be collected, generate artificial samples E.g image flips, rotations, rescaling)
4. Feature selection: Choose the most important features to reduce complexity and avoid fitting noise
5. L1/L2 regularization: Add a penalty term to shrink coefficients (L1 some become zero, L2 values shrink but don't reach zero)
6. Reduce model complexity: Remove layers or reduce the number of neurons to make the model simpler
7. Dropout: Randomly ignore some neurons during training to reduce dependency between them

8. Early stopping: Stop training when validation loss starts increasing to prevent memorization

5. Real-World Case Study — Regression in Business

Title: E-commerce Sales Forecasting, Business Case for Linear Regression

Goal: Help the business optimize production, inventory, and marketing strategies

Data:

1. Historical sales data of products
2. Website traffic data
3. Product categories and promotion history

Model: Multiple Linear Regression

Key Results / Insights:

1. it provided accurate forecasts for future sales
2. Important predictors of sales included website traffic, promotions, and product category
3. Business could plan inventory and marketing more efficiently
4. Demonstrated the practical impact of regression analysis in business decision-making

References:

1: https://builtin.com/data-science/regression-machine-learning

2: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/

3: https://medium.com/@data-overload/understanding-polynomial-regression-a-powerful-tool-for-complex-relationships-d2394a898fd6

4: https://www.ibm.com/think/topics/classification-vs-regression

5: https://net-informations.com/ml/mla/poly.htm?utm_source=chatgpt.com

6: https://medium.com/data-science/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d

7: https://shorturl.at/9pKxC

8: https://www.synaptiq.ai/library/e-commerce-sales-forecasting-a-business-case-for-linear-regression?utm_source=chatgpt.com

By: Mohamed Mohamud