Data Mining Design Using CRISP-DM

Jovan L. Thompson

July 10, 2024

Defining CRISP-DM and The Step Cycles

Defining CRISP-DM

This assignment applies the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to examine the relationship between midterm and final exam grades. The goal is to develop a predictive model that can estimate a student's final exam grade based on their midterm performance. The CRISP-DM methodology is widely used and gives a framework for the phases/step-cycles of a data mining assignment. It provides a structured approach for analyzing and obtaining insights from data, and ensures that the process is repeatable and systematic. The methodology includes six major phases/step-cycles: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phases/step-cycle gives detailed tasks and activities designed to achieve the objectives of the data mining assignment (Luna 2021).

Defining The CRISP-DM Step Cycles

1. **Business Understanding:** This opening phase focuses on understanding the project objectives and requirements from a business perspective. It involves identifying the problem, defining project goals, and formulating a data mining plan. This phase ensures that the business problem drives the project, rather than the data science aspect, by determining business objectives, assessing the situation, establishing analytic goals, and producing a project plan.
2. **Data Understanding:** In this phase, data is collected and examined to understand its characteristics through initial data exploration, data quality verification, and uncovering data insights. This involves collecting, describing, exploring, and verifying the quality of the available data. Initial data collection and familiarization are essential to uncover initial insights and detect data quality issues, ensuring that the data is suitable for subsequent analysis. This phase may involve identifying missing values, outliers, and inconsistencies, as well as gaining an understanding of the data's distribution, relationships, and patterns. Properly understanding the data at this stage is crucial for the success of the entire data mining project, as it lays the foundation for accurate modeling and analysis.
3. **Data Preparation:** This phase includes preprocessing the data to prepare it for modeling by selecting, cleaning, constructing, integrating, and formatting the data

needed. Transforming raw data into a final dataset suitable for modeling is crucial. Tasks in this phase include data cleaning to handle missing values, outliers, and inconsistencies; data transformation to normalize or scale data as needed; data integration to combine data from various sources; and data formatting to ensure the dataset meets the requirements of the chosen modeling techniques. Proper data preparation enhances the quality and relevance of the data, enabling more accurate and robust models. This phase ensures that the dataset is refined and optimized, facilitating effective analysis and predictive modeling in subsequent stages (Luna 2021).

4. **Modeling:** Various modeling techniques are explored, selected and applied to the prepared data during the modeling phase. This phase requires iterative experimentation to identify and refine the most effective model. The models are tuned and calibrated by adjusting their parameters to optimal values to ensure they meet the project goals. This process includes selecting appropriate algorithms, fitting the models to the data, and evaluating their performance using relevant metrics. The goal is to identify the model that best captures the underlying patterns and provides the most accurate predictions or insights. Iteration is key in this phase, as it allows for continuous improvement and refinement of the models to achieve the highest possible performance.
5. **Evaluation:** This phase involves thoroughly evaluating the modeling to ensure they are achieving the desired results and meeting the project and business goals and objectives. If necessary, the models are iteratively refined and re-evaluated to enhance their accuracy and effectiveness. This phase also involves reviewing the entire modeling process to determine the next steps and ensure that the model sufficiently addresses the business problem. Metrics such as accuracy, precision, recall, and other relevant performance indicators are used to evaluate the model. Feedback from stakeholders could also be added to ensure the model's outputs are practical and actionable. The overall objective is to validate that the model not only performs well technically but also delivers value from a business perspective, guiding informed decision-making.
6. **Deployment:** The last phase of CRISP-DM involves deploying the models into a live production environment where they can be used to make informed decisions. This phase includes the initial deployment and the continuous monitoring of the models to confirm they continue to perform well over time. Deployment can involve generating reports, creating dashboards, or implementing automated and repeatable data mining processes that will be used regularly within the organization. Additionally, it involves establishing procedures for maintaining and updating the models as new data becomes available, ensuring their continued relevance and accuracy. Effective deployment

transforms the data mining results into practical tools that drive business value and support strategic decision-making.

Identifying Business Understandings, The Data, and Data Preparation

Identifying Business Understandings

The goal of this assignment is to use the CRISP-DM methodology to develop a least squares regression model which can predict a student's final exam grade based on their midterm exam grade. This model could most directly and obviously be used in real-world situations to assist educators in identifying students who may need additional support to improve their performance on an upcoming final exam. Other uses are gauging overall course performance based on early assessment results, and informing instructional strategies by understanding the relationship between midterm and final exam grades (Hyndman & Athanasopoulos, 2021).

Identifying The Data

The dataset includes the midterm and final exam scores of students. Midterm exam scores is the independent variable X, and final exam scores is the dependent variable Y. There are 12 values for each test and midterm exam scores range from 33% to 94% while final exam scores range from 49% to 90%. The dataset is shown in the table below:

DSC-570 Midterm and Final Exams Table	
X	Y
Midterm exam scores (%)	Final exam scores (%)
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

Identifying Data Preparation

For CRISP-DM, the data preparation steps that apply are shown below:

1. Data Collection: This step involves gathering all relevant data from various sources.
2. Data Integration: This step involves combining data from various sources.
3. Data Cleaning: This step involves identifying and handling missing, duplicate, outlier, or inconsistent data.
4. Data Formatting: This step ensures the data types are appropriate (e.g., numerical, categorical) and that the data is structured correctly for modeling.
5. Data Transformation: This step involves normalization, scaling, encoding categorical variables, and other transformations to make the data suitable for modeling.
6. Feature Engineering: This step includes creating new features from existing ones to improve model performance.
7. Data Splitting: This step involves dividing the dataset into training, testing, and validation, sets, which is crucial for evaluating the model's performance and ensuring that the model generalizes well to unseen data.

Beyond data collection, for this assignment none of the other data preparation steps are needed for the dataset. No data integration is needed and there are no missing values, outliers, etc. in the dataset that need cleaning. Also, no data formatting, transformation engineering, or splitting is needed. The data is suitable for modeling in its original form.

Model Creation

Model Selection and Justification

For this assignment, a simple linear regression model was chosen because it is suitable for predicting the continuous dependent output variable, final exam grades, based on a single independent input variable, midterm exam grades. Linear regression is appropriate for this dataset because we are interested in understanding the linear relationship between the midterm and final exam grades. This method allows us to quantify the strength and direction of this relationship and make predictions based on the midterm grades. The model was created in Jupyter Notebook (file attached) using the method of least squares (Brownlee, 2016).

Analysis of the Results and Deployment

Results Analysis

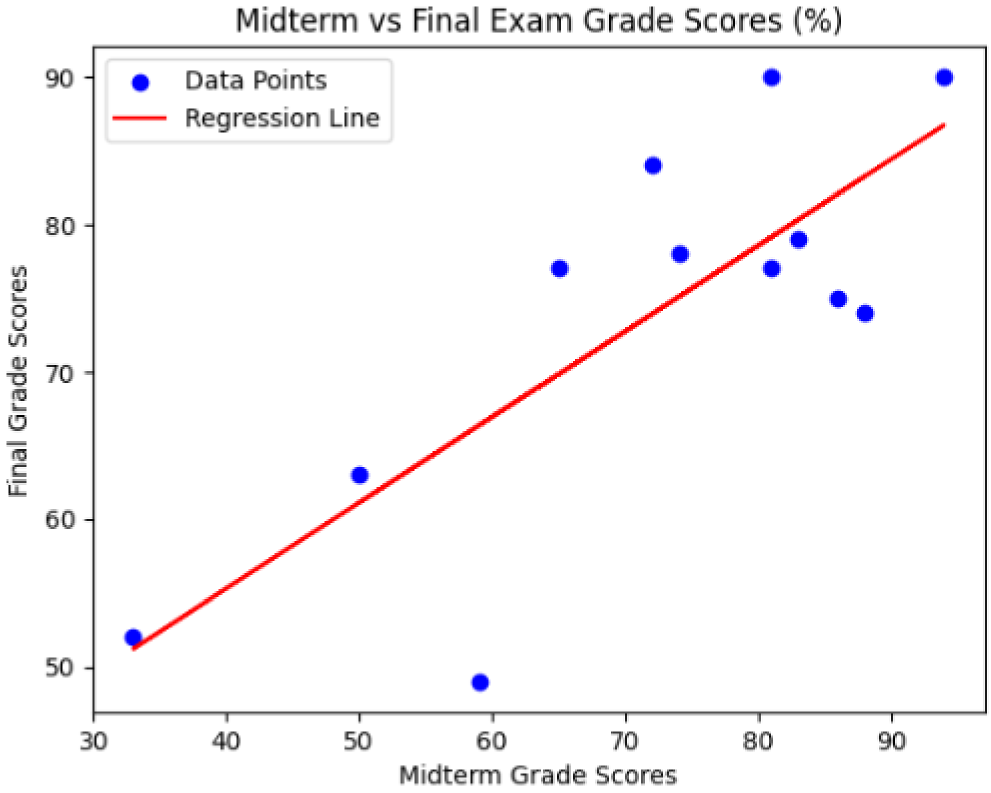
The linear regression model was fitted to the data, and the results were analyzed to assess the model's performance. The model's slope (m) is 0.5816, indicating that for every one-point increase in the midterm grade, the final exam grade is expected to increase by approximately 0.58 points. The intercept (b) is 32.03, which represents the expected final exam grade for a student who scored zero on the midterm exam. The R-squared value is 0.613, meaning approximately 61.3% of the variability in the final exam grades can be explained by the midterm grades. This indicates a moderately strong positive linear relationship between the two variables. The p-value for the slope is approx. 0.003, which is significantly less than 0.05, suggesting that the relationship between midterm grades and final exam grades is statistically significant. Additionally, the model's F-statistic is 15.83 with a p-value of 0.00261, which reinforces the overall significance of the regression model. Lastly, the standard error for the slope is 0.146, and the confidence interval for the slope ranges from 0.256 to 0.907, which provides further insight into the precision of the estimated coefficients. Overall, these metrics indicate that the model provides a reasonably good fit to the data, and captures the general trend between midterm and final exam grades (Hyndman & Athanasopoulos, 2021).

Deployment

The final model was deployed via Jupyter Notebook, which allows for predictions to be made for new unseen data. Given the fitted model's coefficients, predictions can be generated for students' final exam grades based on their midterm grades. As a real-world example, this model could be used in an educational system to provide rapid predictions and insights, helping educators identify students who may need additional support to improve their final exam performance. By monitoring and updating the model periodically with new data, the system can improve and maintain its accuracy and relevance, providing valuable assistance in educational planning and student support.

Plotting Data and Linear Relationship

The data was plotted via Jupyter Notebook to visualize the relationship between the midterm and final exam grades. The key metrics and scatter plot below shows the data points and the fitted regression line: Slope (m): 0.582, Intercept (b): 32.028, R-squared: 0.613, P-value: 0.00261.



The plot indicates a positive linear relationship between the midterm and final exam grades, with the regression line suggesting that for every one-point increase in the midterm grade, the final exam grade increases by approximately 0.582 points.

Using the Method of Least Squares to Find the Equation of the Line

The method of least squares was used to find the equation of the regression line. The formula for the regression line is: y = mx + b where "m" is the slope and "b" is the y-intercept. Using the data obtained from running the model, the equation of the line can be calculated derived from slope and intercept below:

- Slope (m): 0.5816
- Intercept (b): 32.0279
- Equation: y = 0.5816x + 32.0279

This equation represents the linear relationship between given midterm grades (x) and predicted final exam grades (y).

Predicting the Final Exam Grade

In the Jupyter Notebook code using the regression equation, the final exam grade was predicted for a student who received an 86% on the midterm exam. This was verified below where the predicted final exam grade for a student who scored 86% on the midterm exam was calculated as y = 0.5816(86) + 32.0279 = 82.05%.

Conclusion

In this report, I effectively used the CRISP-DM methodology to develop a predictive model for estimating a student's final exam grade based on their midterm grade. Through defining and executing each phase of CRISP-DM, a structured and thorough approach to the analysis was performed. The linear regression model demonstrated a moderately strong positive linear relationship, evidenced by an R-squared value of 0.613 and a statistically significant p-value of 0.00261. Finally, the model's practical application was demonstrated by predicting a final exam grade of 82.05% for a student who scored 86% on the midterm exam. This predictive capability could serve as a valuable tool for educators to identify students needing additional support, which could enhance educational planning and student success.

References

Luna, Z. (2021, August 10). CRISP-DM Phase 3: Data Preparation. Analytics Vidhya. Retrieved from https://medium.com/analytics-vidhya/crisp-dm-phase-3-data-preparation-faf5ee8dc38e

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3nd ed.). Retrieved from https://otexts.com/fpp3/arima.html

Brownlee, J. (2016). A Gentle Introduction to Simple Linear Regression in Machine Learning. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/linear-regression-for-machine-learning/

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining. .Retrieved from https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf