WIL PROJECT

Program Name: Big Data Analytics

Project Code: **CPL-5559-DSMM**

WIL Project Case Study Orientation

Vocational Learning Outcomes (VLOs) Covered in this WIL Project Case Study

- Collect, house, extract, manipulate, maintain, and mine data sets that respond to organizational, financial, or market needs.
- Recommend different systems and network architectures, artificial intelligence, and data storage technologies to support data analytics and Big Data.
- Design and apply data models that meet the needs of a specific operational process or business model.
- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.
- Collaborate effectively with diverse teams to design and present data visualizations that communicate Big Data concepts and information to stakeholders and business professionals.
- Apply business analytics, business intelligence tools and research to support evidence-based decision-making that helps an organization meet their stated objectives.
- Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.
- Implement data security solutions in compliance with corporate security policies, ethical standards, and industry regulations.
- Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.
- Develop artificial intelligence solutions to support administration, decision-making, planning, risk management, logistics, manufacturing, smart devices, and robotics.

Essential Employability Skills (EESs) Covered in this WIL Project Case Study

- Communication
    - It helps to communicate clearly, correctly, and concisely in different forms. These include oral, written, and visual.
- Numeracy
    - This skill set helps to solve mathematical operations effectively with accurate precision.
- Critical thinking & problem solving
    - It is a systemic approach to attempting to resolve problems by analyzing the pros and cons of a decision.

- Information management
    - It helps to locate, select, organize, and document information with the use of technology by analyzing aspects and gathering information from a variety of sources.
- Interpersonal Skills
    - This skill set is important as it helps to respect others' opinions or input. It helps to build teams and maintain relationships to achieve overall team or organizational goals.
- Personal Skills
    - These soft skills are important in developing employability talents, such as dependability, adaptability, and problem-solving skills.

Week 1

This Week's Detailed Case study Information

Congratulations you have been hired as a Data Engineer for Carsy, an automobile consulting company located in Toronto ON.

Carsy Consulting is a leading provider of data-driven insights into the automobile industry. They specialize in helping automobile manufacturers, dealerships, and suppliers make better-informed business decisions by leveraging the power of data. Their team of experienced data analysts, data scientists, and data engineers work together to collect, process, and analyze data from a variety of sources, including sales data, production data, and supply chain data.

The team uses a combination of statistical analysis, machine learning techniques, and data visualization to deliver actionable insights to clients. They have a deep understanding of the automobile industry and use this knowledge to identify key trends and patterns in the data. This allows us to provide their clients with valuable insights that can be used to improve their businesses.

The data infrastructure includes a state-of-the-art data lake, a set of data pipelines for extracting and transforming data, and a set of reporting and visualization tools. They maintain and expand this infrastructure as needed to support our consulting projects. The data lake serves as a central repository for all of the clients' data, which allows the team to easily access and analyze data from multiple sources. The data pipelines are used to extract data from these sources and transform it into a format that can be easily analyzed. The reporting and visualization tools are used to present the findings to clients in an easily understandable format.

Carsy, a consulting firm that specializes in the automobile industry, has been enlisted by Hayai Auto to assist in their expansion into the US market. Hayai Auto, a Japanese automobile company, wants to understand the key factors that influence car pricing in the American market in order to effectively compete with American and European automakers. By utilizing Carsy's expertise in data analysis and their understanding of market trends, Hayai Auto hopes to determine which variables are most important in predicting car prices in the US and how well these variables align with the current market. As a member of Carsy's team, you will play a crucial role in helping Hayai Auto achieve its goal.

- Valeria Bell (Lead Data Scientist)
- Joseph Costa (Data Analyst)
- Louis Berg (Data Engineer)
- Emilee Morrison (ETL Developer)
- Hector Reynolds (Data Manager)

You will be working closely as Data Engineer with these team members.
Valeria Bell is a highly experienced data scientist, with over 8 years of experience in the field. She has a strong background in statistical analysis, machine learning, and data visualization. She is an expert in using these techniques to extract insights from large and complex datasets. Valeria is a confident and decisive leader, who is able to guide her team towards a common goal. She is also known for her ability to explain complex technical concepts in simple terms, making her a valuable asset to her clients. Joseph Costa is a skilled data analyst, who has a strong background in data visualization, statistical analysis, and machine learning. He has experience working with a variety of data sources and has a keen eye for detail. He is an excellent communicator and is able to explain complex data concepts to non-technical stakeholders. Louis Berg is a talented data engineer, with many years of experience in designing, building and maintaining data pipelines. He has extensive knowledge of big data technologies and is able to work with large and complex data sets. He is a problem solver and is always looking for ways to improve the efficiency of the data infrastructure. Emilee Morrison is an experienced ETL developer, with a background in data modeling, data warehousing and business intelligence. She has a strong understanding of data management and governance and is able to build robust, efficient and scalable data pipelines. She is known for her ability to work under pressure and meet tight deadlines. Hector Reynolds is a data manager, with a background in data governance, data quality and data security. He has a deep understanding of data management best practices and is able to implement them in a variety of different contexts. He is a strong communicator and is able to build strong relationships with both technical and non-technical stakeholders.

Week 1 Onboarding Expectations and Participation

Your task this week is to participate in training and orientation for Carsy. You will participate in a variety of exercises that are designed to get to know you better and understand your role within the team. You will participate in Team-building exercises that prepare you for success within the Project. As with any position, you will have an excellent opportunity to build on your skills as a leader so long as you put forth your best effort. Use this week to develop a communication plan with your team and be ready to dive into the deliverables starting next week.

Note: You can make any assumptions that are deemed necessary for each case on a week-by-week basis. You will not be provided direct answers or 100% of the information necessary to complete each deliverable. Instead, focus on delivering the highest quality outcome possible to highlight your talent as a group. You would be presenting these deliverables to Emilee and would want to ensure that the work is of the highest quality.

This section will be available to you for the entirety of the project. However, each subsequent week's case study information may only be available for that week. Be sure to download and save this week's information for future use.

Week 2

Applicable VLOs or EESs for This Week's Case Study

- Collect, house, extract, manipulate, maintain and mine data sets that respond to organizational, financial, or market needs.

This Week's Detailed Case Study Information

Carsy's team will also collaborate closely with Hayai Auto's team to ensure that the project is aligned with their goals and that the final model meets their needs. Additionally, Carsy's team will use their expertise in the automobile industry to provide valuable insights and recommendations throughout the project. The team will also use their data-driven approach to make sure that the model is based on sound statistical principles and that it is robust and reliable. Carsy team is excited to work on this project and help Hayai Auto succeed in the American market.

The project team has been assembled and is ready to begin work on the car price prediction model. As mentioned in week 1 the team consists of data analysts, data scientists, and data engineers who will work together to collect, process, and analyze data from a variety of sources. The project sponsor, a car manufacturer, has provided the team with a list of features that they believe may be relevant for predicting car prices, including the make and model of the car, the year it was manufactured, the mileage, and the location where it is being sold.

The team's first task is to identify the data sources that will be used to obtain these features and assess their quality and relevance. They will also need to set up a data storage solution to store the data that will be used in the model. This might involve setting up a data lake or a relational database, depending on the needs of the project.

In addition to these tasks, the team will also need to set up a version control system to organize the project code and documentation and develop a project plan that outlines the timeline and milestones for the project. The project plan will include the data preparation, model development, and evaluation phases, as well as any other tasks that are necessary to complete the project successfully.

Deliverables for This Week's Case Study

1. Set up a version control system (e.g. Git) and create a project folder structure to organize the project code and documentation.
2. Develop a project plan that outlines the timeline and milestones for the project, including the data preparation, model development, and evaluation phases.

Week 3

Applicable VLOs or EESs for This Week's Case Study

- Design and apply data models that meet the needs of a specific operational process or business model.
- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

You've set up the version control and project plan. This week you need to make sure that you have figured out the storage for the datasets. There are multiple options available to store data and you need to choose the one that works best for you.

This week, the team will be focusing on finding the best storage solution for the datasets. You will be evaluating different options, such as using a data lake or a relational database, to ensure that the team chooses the one that best fits the project's needs. The team will also be looking into cloud-based storage solutions, such as AWS S3 or Azure Blob Storage, to ensure that data is easily accessible and secure. Once the best storage solution is determined, you will be able to move forward with basic exploratory data analysis, including descriptive statistics and understanding the nature of variables. This will provide us with a deeper understanding of the data and will help inform our decisions for the next steps in the project.

Emilee suggested that you may want to create visualizations to gain insights about the data and the relationship between variables such as scatter plots, box plots and histograms. These visualizations can help you identify patterns and trends in the data that may not be apparent from the descriptive statistics alone.

Deliverables for This Week's Case Study

1. Evaluate different storage solutions for the datasets, including data lake, relational database, and cloud-based storage options such as AWS S3 or Azure Blob Storage.
2. Choose the best storage solution that fits the project's needs.
3. Perform basic exploratory data analysis, including descriptive statistics and understanding the nature of variables.
4. Create visualizations, such as scatter plots, box plots and histograms, to gain insights about the data and the relationship between variables.
5. Use the results of the exploratory data analysis to inform decisions for the next steps in the project.

Week 4

Applicable VLOs or EESs for This Week's Case Study

- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

The team has made significant progress in designing the infrastructure for the project and is now turning its attention to data security and privacy. Data security and privacy are critical concerns in any project, and this one is no exception. Valeria, a security expert on the team, has reminded us that we need to take steps to identify and mitigate any risks and concerns related to data security, privacy, and breaches.

This week, the team will be working together to identify any potential risks and concerns related to data security and privacy. We will be assessing the data storage solutions we have chosen to ensure that they meet industry standards for security and privacy. We will also be evaluating the data handling procedures we have in place to ensure that they comply with regulations and best practices for data security and privacy.

In addition to identifying risks and concerns, the team will also be taking steps to mitigate them. This may include implementing additional security measures, such as encryption, access controls, and monitoring systems. The team will also be developing incident response plans to ensure that we are prepared to respond to any data breaches or other security incidents.

As part of this, the team will also be conducting a risk assessment to identify the likelihood and impact of different types of security breaches and create a plan to mitigate them. Moreover, the team will be educating themselves and the stakeholders about the data security and privacy protocols and best practices, to ensure that everyone on the project is aware of their responsibilities and the importance of protecting data.

Deliverables for This Week's Case Study

1. Collaborate with the team to identify potential data security, privacy, and breach risks associated with the project
2. Develop a plan to mitigate identified risks and concerns
3. Review and update data handling and storage protocols to ensure compliance with industry standards and regulations
4. Provide regular updates to the project sponsor on the team's progress in addressing data security and privacy concerns

Week 5

Applicable VLOs or EESs for This Week's Case Study

- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

With the recent news that the project is going to continue for the long run, the team needs to focus on automating the process of extracting and transforming data for the model-building process. Valeria has emphasized the importance of using data pipelines to achieve this automation.

This week, the team will be researching different options for building data pipelines. This includes looking into popular tools such as Apache Kafka, Apache NiFi, and AWS Glue. The team will also be evaluating the pros and cons of each option to determine which one best fits the project's needs. This will ensure that data can be extracted and transformed in a consistent and efficient manner, allowing the model to be updated with new data in a timely fashion.

Additionally, Emilee suggested that the team also consider implementing a monitoring and alert system to ensure that the data pipeline is running smoothly. This will help the team identify and resolve any issues that may arise and ensure that the data is being processed correctly.

Overall, this week's task is to research and implement a data pipeline solution that will automate the process of extracting and transforming data for the model-building process, while also ensuring that the pipeline is running smoothly and efficiently.

Deliverables for This Week's Case Study

1. Begin to Implement the chosen data pipeline solution to automate the process of extracting and transforming data for the model-building process.
2. Test and evaluate the effectiveness of the implemented data pipeline, making any necessary adjustments to ensure it is working as expected.
3. Document the process and results of the implementation for future reference and to share with the team.
4. Be mindful of the data security, data privacy and data breach risks while implementing the pipeline. Collaborate with the rest of the team to identify and mitigate these risks and concerns.

Week 6 – Midterm Week

Mid-Term Panel Evaluation Preparation

The team will prepare for the Mid-Term Panel Evaluation this week. For the Team Presentation create a professional multimedia presentation highlighting the key aspects of your project thus far. Please see Moodle for full details.

Presentation

CONTENT

- Overview of work in the Iconic
- Highlight three key areas you find of interest:
  - Two areas related to weekly work completed
  - One area to highlight PD or other activity
- Apply reflecting skills
- Present the importance/benefit of work to Iconic.

FORMAT & LAYOUT

- Research and design professional presentations with a popular tool such as PowerPoint, Canva or Adobe, free of spelling and grammatical errors. Images and other graphics used should represent the company and communicate the brand's color scheme.
- Team effort: every team member needs to present, contribute and account for the contribution towards the presentation

Week 7

Applicable VLOs or EESs for This Week's Case Study

- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

Everybody has appreciated the work you've done so far. Well done!

With the design phase of the project now complete, the team is ready to begin the model-building process. You have been provided with a dataset containing both independent and dependent variables, and the goal is to construct a model that can accurately predict the price based on the independent variables. This is a supervised learning problem, as the objective is to use the data to train the model to make predictions.

In order to build the model, you will need to perform various tasks such as data pre-processing, feature selection, and model selection. This will involve cleaning the data, handling missing values, and normalizing the features. You will also need to select the appropriate algorithm that fits the problem and the data, and train and evaluate the model using different metrics.

Emilee has asked you to evaluate the model performance and fine-tune it as needed by trying different combinations of parameters, and ensembling different models. This process may require multiple iterations, so it is important to keep track of the progress and document the findings.

As the project is a long-term one, you will also need to consider the scalability and maintainability of the model, and implement a strategy to monitor and retrain the model as new data becomes available.

Deliverables for This Week's Case Study

1. Select the most appropriate algorithm or combination of algorithms to use for the model-building process.
2. Prepare the dataset for model-building by cleaning and pre-processing the data, as needed.
3. Train and validate the model using the prepared dataset.
4. Evaluate the performance of the model and make adjustments, as needed.
5. Document the entire model-building process and results, including any challenges faced and solutions implemented.

Week 8

Applicable VLOs or EESs for This Week's Case Study

- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.

This Week's Detailed Case Study Information

In this week, you and Valeria will be focusing on evaluating and validating the model that was developed in the previous week. This will involve experimenting with different algorithms and features to determine the impact they have on the predicted prices. As part of this process, Valeria has suggested going through one more iteration of feature engineering to ensure that the features selected are the most relevant and useful for the model. Additionally, the team will be splitting the dataset into training, test, and validation sets in order to evaluate the model's performance on unseen data. This will help to ensure that the model is robust and can generalize well to new data. Overall, the goal for this week is to fine-tune the model and make any necessary adjustments to improve its accuracy and performance.

Emilee has suggested you use a variety of tools and techniques to evaluate the model, including metrics like mean squared error and R-squared. You will also test the model on new data to see how well it generalizes to unseen situations.

In addition to evaluating the model, you should also be communicating the results and recommendations to stakeholders, such as management or clients. This may involve preparing presentations or reports to explain the model's performance and any areas for improvement.

Deliverables for This Week's Case Study

1. Research and evaluate different algorithms and feature sets to determine the impact on model performance
2. Conduct an additional iteration of feature engineering to identify and add any additional relevant features
3. Split the dataset into training, test, and validation sets
4. Evaluate the model's performance on the unseen validation set and make any necessary adjustments to improve performance.

Week 9

Applicable VLOs or EESs for This Week's Case Study

- Implement data security solutions in compliance with corporate security policies, ethical standards, and industry regulations.
- Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.

This Week's Detailed Case Study Information

You have experimented with different features in the previous week now it's time to experiment with the different hyperparameters to improve the performance of the model. Adjusting the hyperparameters with the aim to improve the accuracy of the model is commonly called hyperparameter tuning.

This week you will be collaborating with Hector and Joseph to fine-tune the hyperparameters of the model developed in the previous week. The goal is to improve the accuracy of the model by adjusting the hyperparameters. This process is commonly referred to as hyperparameter tuning. There are several methods that can be used to perform hyperparameter tuning such as grid search, random search, and Bayesian optimization. Together with Hector and Joseph, you will research these options and decide on the best approach for the current project. Emilee asked you to develop the code to implement the chosen approach, test it and evaluate the results.

You will need to continue communicating their results and progress to stakeholders, such as management or clients. This may involve preparing presentations or reports to explain the model's performance and any changes that have been made. Overall, you will continue to refine and improve the model that was built earlier. You can always refer to the documentation and review similar projects done in the past when in doubt or confusion.

Deliverables for This Week's Case Study

1. Research different options for performing hyperparameter tuning.
2. Implement a chosen method for hyperparameter tuning on the model.
3. Evaluate the performance of the model with the fine-tuned hyperparameters.
4. Compare the performance of the model before and after hyperparameter tuning.
5. Document the process and results of hyperparameter tuning for future reference.
6. Document the model in detail, including the steps that were taken to build it, the assumptions that were made, and any limitations. This documentation will be important for maintaining and updating the model in the future.

Week 10

Applicable VLOs or EESs for This Week's Case Study

- Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.
- Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.

This Week's Detailed Case Study Information

When working on a machine learning project, it is common practice to train multiple models on the dataset and select the one that performs the best. However, there is always the potential for improvement, as it is not certain that the chosen model is the optimal solution for the problem at hand. To try and improve the model, it is possible to use neural networks, which are known for their effectiveness when dealing with large volumes of data. However, it is also worth considering using neural networks in situations where the data is not as extensive, as they may still provide good performance. Ultimately, it is important to evaluate the performance of different models and choose the one that works best for the specific problem.

Valeria has implemented neural networks to solve different kinds of regression and classification problems and she is curious if the Neural Network will improve the performance in this case. Some of the widely used types of Neural networks are Deep Neural Networks, wide Neural Networks and Deep and wide neural Networks. Valeria requested you go through the documentation of Keras and Tensorflow and implement the neural network and observe if they perform better than the model developed in the earlier weeks. This week, you need to

- List out the differences and use cases of Deep and Deep and wide Neural Networks.
- Make a report comparing the performance of the conventional machine learning model and Neural Networks. And conclude if it makes sense to use NN in this use case

Deliverables for This Week's Case Study

1. Use any two Neural Network models to predict the prices of cars.
2. Compare the performance of your Neural Network with the conventional model developed in earlier weeks
3. Analyze the complexity and speed of your model and NN model and log it.

Week 11

Applicable VLOs or EESs for This Week's Case Study

- Develop software applications, algorithms, and artificial intelligence models to manipulate, correlate and reduce data sets and produce project documentation and reports.
- Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.

This Week's Detailed Case Study Information

It is clear from management's decision that the project is going to stay for long run. That's why you've created the data pipeline to ingest the data in the previous weeks.

This week you will be focusing on breaking down the monolithic architecture of the code and creating modular and reusable components. This will ensure that the code is easy to maintain and update as new data is ingested in the future. You will be working closely with the team to identify areas of the code that can be refactored into reusable components and implementing these changes. Additionally, you will be documenting the code and creating a clear structure that can be easily understood by other team members. This will help in efficient collaboration and make it easy to add new features and functionalities in the future.

Additionally, the team is planning to integrate this built model with other internal system using REST API, and Emilee has asked you to come up with architecture where the model is your core business logic and there is an API interface to interact with the user queries.

Overall, the goal for this week is to create a modular and reusable code base that can be easily maintained and updated as the project continues to evolve.

Deliverables for This Week's Case Study

1. Break down the monolithic code architecture into manageable modules and reusable components
2. Identify areas of the code that can be turned into modular and reusable components
3. Refactor the code to create modular and reusable components
4. Test the modular and reusable components to ensure they are working as expected
5. Document the new modular and reusable components and update project documentation accordingly.
6. Develop an architecture diagram where the build model can be used via REST API.

Week 12

Applicable VLOs or EESs for This Week's Case Study

- Develop artificial intelligence solutions to support administration, decision-making, planning, risk management, logistics, manufacturing, smart devices, and robotics.

This Week's Detailed Case Study Information

This week you will be focusing on finalizing the model and preparing it for deployment. The team has spent the previous five weeks exploring the data, identifying relevant features, building a model, and evaluating and refining its performance.

The team is focused on completing any final adjustments to the model and preparing it for use by others. This may involve testing the model on new data to ensure it generalizes well, documenting the model in detail, and creating an interface or application to allow users to access it.

You will also need to communicate the final results of the project to stakeholders, such as management or clients. This may involve preparing presentations or reports to explain the model's performance and any recommendations for further improvements or next steps. Overall, the goal for this week is to finalize the model and prepare it for deployment, and to communicate the results of the project to stakeholders.

Deliverables for This Week's Case Study

1. Continue making adjustments to the model in order to improve its performance. This could involve testing different hyperparameter settings, adding or removing features, or using different modeling techniques.
2. Validate the model by testing it on a separate dataset that has not been used in the training process. This can help determine how well the model generalizes to new data.
3. Document the model in detail, including the steps that were taken to build it, the assumptions that were made, and any limitations. This documentation will be important for maintaining and updating the model in the future.
4. Prepare the model for deployment, which may involve converting it to a format that can be easily used by others or creating an interface or application to allow users to access the model.
5. Communicate the final results of the project to stakeholders, including any recommendations for further improvements or next steps.

Week 13

Applicable VLOs or EES for This Week's Case Study

- Identify and assess data analytics and Big Data business strategies and workflows to respond to new opportunities or provide project solutions.
- Deliver data-oriented projects using data science, business analysis, and project management principles, tools, and techniques.

This Week's Detailed Case Study Information

Submission of Project Report + Practice Presentation

- Finish Project Report for submission your final submission is due this week. Be proud of the work you have completed in this project, now you can spend time polishing your presentation and making sure you will capture the stakeholder's attention in a positive way.
- Review APA Guidelines and ensure your project has followed them. This is particularly important.

Hone your presentation skills.

- A Presentation for your Car price prediction project is meant to highlight your research findings and the conclusions, opportunities, and best practices that you can be followed on other projects. The analysis of your findings is one of the most important parts and should be conveyed in your presentation.

Deliverables for This Week's Case Study

1. Final Project Report – this is your final document with all supporting resources: including any appendices. Bibliography and reference in APA format required.
2. Feedback Video
    - Prepare to answer questions regarding the project on client expectations, Job Market, and on how you will sell your product

Week 14

Preparing for Your Final Week Activities

It is the end of your work term. Your supervisor is grateful for your efforts. The final week contains activities which include both individual and teamwork efforts. Take this opportunity to shine bright in the final activities.

Final Week Deliverables and Format Requirements

Your supervisor will provide you with more detail about the Final Week responsibilities.

WIL Project Completion

Following completion of the Final Week activities, you will be notified by your supervisor if you pass or fail the WIL Project.

Appendix

Acronym Used

HDFS: Hadoop Distributed File system
PCA: Principal Component Analysis
NLP: Natural Language processing
PD: Personal Development
HQL: Hive query Language
SQL: Structured Query Language
RBAC: Role Based Access Control
ML: Machine Learning
DE: Data Engineer
PM: Project Manager
PII: Personal identifiable information
AI: Artificial Intelligence
MSE: Mean Square Error
MAE: Mean absolute Error