**BPC Work Plan**

Executive Summary:

As the Data Engineering Manager, my proposed approach for BPC's project is to implement the ChemBERTa algorithm in their workflow for chemical fingerprinting used to represent molecules for property prediction in downstream applications. The approach will be modular, divided into phases, with clearly defined deliverables, skill sets required, potential outcomes, and milestones.

Phase 1: Data Preparation and Exploration The first phase involves preparing and exploring the data required for the project. This phase will require data scientists with expertise in chemistry and data engineering. The expected deliverables in this phase are clean and well-structured data, a comprehensive data dictionary, and data exploration reports. Potential outcomes in this phase are identifying patterns in the data that can influence downstream applications.

Phase 2: Algorithm Implementation The second phase involves the implementation of the ChemBERTa algorithm. This phase will require data scientists with expertise in natural language processing, deep learning, and cloud computing. The expected deliverables in this phase are a well-implemented algorithm, a trained model, and validation reports. Potential outcomes in this phase are improved accuracy in predicting molecular properties.

Phase 3: Deployment and Integration The third phase involves deploying the algorithm in a cloud environment and integrating it into BPC's workflow. This phase will require data engineers with expertise in cloud computing and API development. The expected deliverables in this phase are a scalable and reliable API, documentation, and training materials. Potential outcomes in this phase are an efficient and user-friendly workflow for molecular property prediction.

Phase 4: Documentation and Maintenance The fourth phase involves documenting the implemented solution and maintaining it for future use. This phase will require data engineers with expertise in documentation and maintenance. The expected deliverables in this phase are well-documented solution, maintenance plan, and support. Potential outcomes in this phase are improved solution sustainability and reduced downtime due to maintenance issues.

A rough timeline breakdown and milestones for completing this project in 7 months:

Milestones:

* Data Preparation and Exploration: 4 weeks
  + Identify data sources and acquire data
  + Perform data cleaning and pre-processing
  + Transform data into a suitable format for training
  + Conduct exploratory data analysis (EDA) to gain insights into the data
* Algorithm Implementation: 12 weeks
  + Train the ChemBERTa algorithm using pre-processed data
  + Fine-tune the model as necessary
  + Validate the model using a test set
* Deployment and Integration: 8 weeks
  + Deploy the trained model on a cloud-based platform
  + Integrate the model into an API for predictions
  + Test the API to ensure it is working correctly
* Documentation and Maintenance: 4 weeks
  + Document the entire workflow, including data sources, pre-processing, training, deployment, and API integration
  + Provide support to the client to ensure the model continues to function as expected

Potential outcomes:

* Improved accuracy in predicting molecular properties
* Efficient and user-friendly workflow for molecular property prediction
* Increased revenue and cost savings for BPC

Resources & Skillsets

Estimated team size to complete for this project in 6 months:

* 1 Project Manager to oversee the project and ensure timely delivery of each phase
  + Experience managing complex data projects and be familiar with project management methodologies, such as Agile or Waterfall. Developing project plans, identifying and managing risks, and tracking progress to ensure timely delivery.
  + Excellent communication skills to collaborate effectively with the team, stakeholders, and clients. Having the ability to convey technical concepts to non-technical stakeholders and communicate project status and risks effectively.
  + Does not need to be an expert in molecular property prediction, they should have a good understanding of the technology and concepts involved in the project (including understanding of data pre-processing, machine learning algorithms, cloud infrastructure, and API development)
  + Able to identify and solve problems effectively, either by providing solutions or escalating issues to higher management.
  + Have excellent time management skills to ensure timely delivery of each phase of the project.
* 2 Data Scientists with expertise in molecular property prediction, deep learning, and data pre-processing (At least one being BPC Data Scientist)
  + Have a deep understanding of deep learning algorithms and techniques. Specifically Neural Networks. Graph Neural Networks (GNN) preferred.
  + Experienced in data cleaning, pre-processing, feature engineering, and data transformation to ensure the data is in a suitable format for training.
  + Preferably they would have a good understanding of the chemistry behind molecular property prediction, including molecular descriptors and fingerprints, and quantum chemistry.
    - A small likelihood of having one on staff which is why we recommend having at least 1 BPC Data Scientist.
* 1 Software Engineer to develop and maintain the API for predictions. Or 1 Software Engineer and 1 Data Engineer at 50% capacity each.
  + Experience in cloud infrastructure, such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP), to deploy the model in a scalable and reliable way.
  + Experience in designing and developing APIs using RESTful architecture and API development frameworks, such as Flask or Django.
  + Proficient in programming languages such as Python or Java and have experience in developing backend services to support the API.
  + Experience in database management, including designing and developing database schemas, SQL queries, and data storage.
  + Experience in software testing, including unit testing, integration testing, and performance testing, to ensure the API functions correctly.
  + Experience in DevOps practices, including continuous integration and deployment (CI/CD), containerization, and infrastructure as code (IaC).
* 1 Cloud Infrastructure Engineer to manage the cloud-based infrastructure for training and deployment
  + Expert in cloud infrastructure, such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP), and be able to set up and manage cloud resources such as virtual machines, load balancers, and storage.
  + Experience in setting up and configuring network security, virtual private networks, security groups, and access control lists (ACLs), to ensure the cloud infrastructure is secure.
  + Experience in containerization technologies such as Docker and Kubernetes to deploy the model and API in a scalable and reliable way.
  + Experience in infrastructure as code (IaC) tools such as Terraform or CloudFormation to manage the cloud infrastructure using code.
  + Experience in setting up monitoring and logging tools, such as CloudWatch, Data Dog, Prometheus, or Grafana, to monitor the health and performance of the cloud infrastructure.

Costs: The costs will depend on the expertise required and the duration of each phase. The expected costs include data acquisition and preparation costs, cloud computing costs, and salaries for data scientists and data engineers. A detailed cost estimate will be provided after scoping the project.

Architectural Diagram: The architectural diagram will be provided after scoping the project.

Gantt Chart: The Gantt chart will be provided after scoping the project.

Note that these timelines are rough estimates, and the actual timeline may vary based on the complexity of the project and the team's level of expertise. Also, the timeline may need to be adjusted based on the specific requirements of the project and the available resources.

Sent from my iPhone