**Prathamesh Joshi**

Ann Arbor, Michigan • prathuj@umich.edu • +1 (734) 596-8174 • [LinkedIn](https://www.linkedin.com/in/prathamesh0902) • [GitHub](https://github.com/prathamesh0902/) • [Kaggle](https://www.kaggle.com/learn/prathamesh0902)

**EDUCATION**

|  |  |
| --- | --- |
| University of Michigan - Ann Arbor, Michigan | 2022 - 2024 |
| M.S. in Data Science | GPA: 3.42 |
| Relevant coursework: Machine Learning, Data Mining, Data Manipulation, Data Visualization, Statistical Theory | |
| Indian Institute of Technology (IIT BHU) - Varanasi, India | 2014 - 2019 |
| B.Tech. and M.Tech. in Industrial Chemistry | GPA: 3.68 |

|  |
| --- |
| What is the date you can start the job?  The start date on my Employment Authorization Document (EAD) is August 2nd and I can start my job immediately  What is your expected Salary?  70k $ to 90k$ annually  33$ / hr - 43 $ / hr  Where do you see yourself in next 5 years?  Career paths, mentoring, training courses  build your own career path  project management, leadership |

**SKILLS**

|  |  |
| --- | --- |
| **Technical:** | Python (Tensorflow, PyTorch), SQL, R, Tableau, Power BI, Microsoft Excel, SQL |
| **Functional:** | Azure Synapse Analytics, Azure Data Factory, ETL Development, Talend |

**CERTIFICATES**

|  |  |  |
| --- | --- | --- |
| **Microsoft Azure:** | Azure Fundamentals [Link](https://learn.microsoft.com/en-us/users/prathamesh0902/credentials/8c030d4094958a47) , Azure Data Fundamentals [Link](https://learn.microsoft.com/en-us/users/prathamesh0902/credentials/e7a93bba218501ae) | |
| **LinkedIn Learn:** | Database Administrator, Salesforce for Admins, Skills for Healthcare Data Analysis | |
| **Kaggle:** | Machine Learning Explainability, Feature Engineering, Data Visualization | |
| Tell me about yourself / work experience  I am Prathamesh Joshi. I have more than 5 years of experience in Data Science field particularly in Data Science, Machine Learning and Reporting roles. I have recently completed my Masters in Data Science from University of Michigan Ann Arbor. Previously I have an experience of working at Axtria, India. I have 2 internship experiences in Data Analytics field. ~~Post my graduation, I have started my journey of self Learning and improving my skills~~  ~~I am eager to learn new technologies and hence~~ I have completed Azure fundamental certifications and also LinkedIn learning on Healthcare industry. I am expecting to complete a Microsoft Power BI certification in 1 week from now. | | How do you stay updated with the latest trends and technologies in data analysis?  Certifications, Linked Learning, YouTube  Which data analysis tools and software are you proficient in?  Python - multiple projects, seaborn, matplot for plotting  Power BI / Jasper - data visualization tools  Salesforce - CRM tool  Azure - cloud technologies, engaged with recent technologies | |

**WORK EXPERIENCE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Technocolabs Software, Data Science Intern | Mumbai, INDIA** Jun’23 – Aug’23   |  | | --- | | • Executed data pre-processing, feature engineering, and map visualizations leveraging libraries such as caret and ggplot2 in R to predict loan prepayment risk through various customer segments. | | • Developed and implemented a comprehensive data science pipeline by implementing Decision Trees, Random Forest and SVM training and evaluation, **enhancing performance up to 92%** through hyperparameter tuning. | | • Integrated data into Power BI to create 10 dynamic dashboards and PowerPoint decks **reducing report generation time by 20%.** | |
| How have you used Machine Learning models? |
| **Axtria Ingenious Insights, Senior Salesforce & Data Analyst | Gurgaon, INDIA** Jul’19 – Jul’22  • Configured administrative settings, user roles/profiling to manage security of **200+ users on daily basis** through Lightning custom Apps.  • Maintained active communication with stakeholders to resolve 80% of user requests within 24 hours using ServiceNow and JIRA.  • Successfully completed **transition of 3 teams, 30 + custom objects** and deployment of **18 Salesforce modules** utilizing Salesforce tools for custom coding and functionalities like Triggers, workflows, Debugging, AWS S3 integration **improving product and data quality by 20%**.  • Conducted A/B testing and hypothesis testing, UAT for new releases in combination with practicing software development life cycle of Agile frameworks, Scrum and Kanban reducing 5 person hours/ week, while using GitHub for tracking product development  • Generated **100+ reports** with Microsoft Excel, **Tableau, Power BI**, and **Jasper Studio** for visualizing user performance in terms of % target achieved, login history and aggregations in form of hierarchical summary delivered to managers and supervisors. |
| EXCEL: data cleaning, SUMIF, CountIF, Concatrnate: for conditional calculations and text manipulation.  Vlookups, Hlookups for Lookup functions for retrieving data from different ranges.  PivotTables: Essential for dynamic and interactive data summarization when computing login history tally over month. |
| • Integrated and analyzed **500,000+** pharmacy transaction records from Salesforce to identify key customer segments and purchase behaviors, leveraging Python and SQL for data manipulation and cleaning.  • Developed and deployed machine learning models within Salesforce to predict customer purchasing patterns, resulting in a **25% increase** in targeted marketing campaign effectiveness and a **15% rise** in sales conversions.  • Successfully completed **transition of 3 teams, 30 + custom objects** and deployment of **18 Salesforce modules** utilizing Salesforce tools for custom coding and functionalities like Triggers, workflows, Debugging, AWS S3 integration **improving product and data quality by 20%**.  • Conducted A/B testing and hypothesis testing, UAT for new releases in combination with practicing software development life cycle of Agile frameworks, Scrum and Kanban reducing 5 person hours/ week, while using GitHub for tracking product development  • Generated **100+ reports** with Microsoft Excel, **Tableau, Power BI**, and **Jasper Studio** for visualizing user performance in terms of % target achieved, login history and aggregations in form of hierarchical summary delivered to managers and supervisors. |
| Lead data science projects from start to finish  **data collection, processing, analysis, visualization and prediction modelling.** |
| **Indian Institute of Technology (IIT) Mandi, SQL Database Management Intern | INDIA** Apr’17 – Jul’17  • Implemented security protocols for **10+ databases (SQL Server, Oracle DB, PostgreSQL),** ensuring data integrity and managing access permissions for 50 + users.  • Spearheaded data management, including data modeling, data cleaning, query optimization, programming, and indexing, resulting in a **12% reduction in processing time** through proactive project management, streamlining operations and improving business intelligence efficiency. |
| **SQL Tools** |

**ACADEMIC PROJECTS**

|  |  |  |
| --- | --- | --- |
| **Market Insights: Tracking Industry Trends, Market Share and Sales** [Link](https://github.com/prathamesh0902/Academic_Projects/tree/main/Data_Visualization) | | Apr ’24 |
| • Conducted data cleaning and transformation using **Power Query**, achieving a **50% faster reporting** turnaround time with real-time data refresh capabilities.  • Utilized DAX (**Data Analysis Expressions**) to create calculated columns and measures for advanced data analysis and visualization, revealing a **15% growth in sales** volume quarter-over-quarter by tracking and analyzing sales data.  • Developed interactive Power BI dashboards, enabling a **20% improvement** in strategic decision-making efficiency. | | |
| How did you use reporting softwares?  Power BI - role based access  Salesforce Reporting - system logs  Jasper Reporting - parameter based reporting  What experience do you have with data visualization tools? Can you provide examples of the types of visualizations you have created?  Provide goals, sales, % target achieved  Hierarchical summary, Users, Managers  Overview summary page year, monthly details page  Can you walk us through your process for creating a detailed report from raw data?  Data Import,  Data modelling - creating semantic relations / measures - DAX,  Show / hide fields, creating measures  Selecting visuals, Role based permissions  Why data reporting  Fascinated by the way in which data can be represented  Creating presentations have always intrigued me to represent my work achievements - technical stats to non-technical audience  it comprises of data engineering, warehousing, outcome prediction  also get to be a part of Business Intelligence team and learn more about healthcare | Explain your experience in Power BI / Tableau  imported data, role-based dashboards  Power BI was strategically employed post-integration with SQL Server to generate visualizations for customer engagement. Different summary charts were made for different hierarchy of users from End Users, Managers to Executives allowing for dynamic and real-time visualizations.  Dynamic Dashboards  1. Interactivity: Dynamic dashboards allow user interaction; static dashboards do not.  2. Data Connectivity: Dynamic dashboards update with live data; static dashboards use fixed data snapshots.  3. Customization: Dynamic dashboards offer customizable views; static dashboards are unchangeable after creation.  4. User Experience: Dynamic dashboards enhance engagement through interactive elements such as slicers, drill-throughs, and clickable visuals; static dashboards provide a simpler, fixed view.  5. Performance: Dynamic dashboards may be slower due to live data; static dashboards load faster but lack real-time updates. | |
| **Anime Odyssey: Recommendation by Exploring Genres, Clustering & Trends** [Link](https://github.com/prathamesh0902/Academic_Projects/tree/main/Data_Mining) | | Feb ’24 |
| • Utilized predictive models and collaborative filtering for generating personalized anime recommendations resulting in a **20% increase in user engagement**, achieved by employing Jacob similarity and Mutual Information score statistics to reveal trends, patterns, and viewer preferences for content curation by combining the data mining task using Pandas and PyTorch libraries in python and spark.  • Visualized genre-specific word clouds by comparing and interpreting 5+ unique word patterns in anime synopses with NLP. | | |
| Large Datasets   * Data Extraction and Cleaning - python, pandas, 2 million * Data normalization - MinMaxScaler() * select specific set or recommendations based on assumptions - 50 k * Genre hierarchical clustering - based on jacob similarity * Principal component analysis, SVD, Mutual Information   Usage: When dealing with massive datasets that may not fit into memory entirely, chunking allows you to process a portion (or chunk) of the data at a time.  NLP: use batch size for training model | | |
| **Beyond Scoreboards: Cricket Data Analysis and Interpretability of ‘Player of the Match’ Awards** [Link](https://github.com/prathamesh0902/Academic_Projects/tree/main/Data_Manipulation_Analysis) | | Nov ’23 |
| • Conducted extensive Data Visualization for 10 + data tables leveraging Seaborn to include raincloud plots, swarm plots, radar plots etc.  • Developed player clusters utilizing data analytics techniques like PCA, distilling information from the top 5 principal components by **reducing data size by 400%.**  • Applied Machine Learning and Regression analysis from Scikit-learn for **achieving R-squared value up to 92%** and were interpreted by employing SHAP analysis to elucidate important features. | | |
| Data Inconsistencies  In a Beyond Scoreboard project involving the integration of multiple databases,  variations in data formats and recording practices.  comprehensive data mapping and cleansing strategy.  establish standardized data formats,  the combined file matched with online records  maintaining data consistency throughout the project.  Did you create data dictionaries for a database?  Can you describe your experience with Python? Provide examples of projects where you utilized these languages  Python for data visualization, manipulation and analysis. NLP  Used pandas and numpy for data cleaning and data comprehending  scikit learn for PCA, applying machine learning models  Matplotlib and seaborn for data visualization  Extracted data from a webpage | How do you ensure the accuracy and reliability of your Python scripts?  For Python scripts, focus on unit testing by verifying the correctness of individual functions. Leverage version control systems like Git for tracking changes, facilitating identification and resolution of issues. Conduct code reviews to obtain feedback on logic and potential enhancements. Implement robust error handling, comprehensive documentation, and test scripts with diverse data scenarios, including edge cases. Manage dependencies using virtual environments or package managers to ensure a consistent and reliable environment. These practices collectively contribute to the accuracy, reliability, and maintainability of Python scripts throughout development. | |
| **Comparative Analysis of ARIMA & Neural Network Models for forecasting International Flight Departure** [Link](https://github.com/prathamesh0902/Academic_Projects/tree/main/Applied_ML) | | Nov ‘23 |
| • Evaluated ARIMA's predictions against **Deep Learning models** like **LSTM** and **Neural Nets** using TensorFlow for interpreting trends, seasonal patterns and discrepancies to achieve an **8% improvement in forecasting** accuracy with the LSTM model.  • Compared model accuracies for forecasting with a precision of achieving a MAE reduction of 5% and **increased computing speed by 2x**. | | |
| SARIMA extends ARIMA to handle **seasonal** patterns in the data   * **Seasonal AR (AutoRegressive):** Similar to AR, but applied to seasonal lagged observations (e.g., data points from the same season in previous cycles * **Seasonal I (Integrated)**: Seasonal differencing to remove seasonality * **Seasonal MA (Moving Average)**   SARIMAX(p, d, q)(P, D, Q)[m] + X : **X**: Represents the exogenous variables added to the model. | | |
| **Lyrics Generator by LSTM supported Natural Language Processing |** [Link](https://github.com/prathamesh0902/Academic_Projects/tree/main/Natural_Lang_Processing/)   * Developed a creative lyrics generator using LSTM and RNN models, achieving a 75% coherence rate in generated lyrics. * Implemented robust text preprocessing, tokenization, reducing preprocessing time by 40% while maintaining data quality. * Applied NN to create a multiclass classifier, achieving a notable BLUE score of 60% for bigram metric, outperforming baseline models by 25%. | | |
| Have you used any other Python Libraries?  Deep Learning experience | | |

|  |
| --- |
| 1. Databricks  Example: In Databricks, I developed a complex ETL pipeline that processed terabytes of raw log data from multiple sources, transforming it into a structured format for analysis. The query involved multiple stages, including data cleansing, aggregating user sessions, and joining with reference tables to enhance the data with additional context. I leveraged Databricks' Delta Lake for handling incremental updates efficiently and used Spark SQL for complex joins and aggregations, ensuring that the pipeline was both scalable and performant.  2. SQL  Example: I wrote a SQL query that performed a recursive Common Table Expression (CTE) to calculate the hierarchy of employees in an organizational structure. The query needed to dynamically traverse a self-referential table where each employee had a manager, and return a report showing each employee's level in the hierarchy and their respective reporting chain. The query also included several subqueries and window functions to calculate cumulative metrics for each level in the hierarchy.  3. Spark  Example: In Apache Spark, I developed a complex data processing job that involved reading a large dataset from HDFS, applying multiple transformations, and then writing the results back to a distributed storage system. The job included operations like filtering, joining with other large datasets, and aggregating data based on several keys. I optimized the job by using Spark's DataFrame API, implementing partitioning and caching strategies to reduce shuffle operations and improve performance.  4. Python  Example: Using Python, I created a data pipeline that integrated data from Google Sheet APIs, cleaned and transformed the data, and then applied machine learning algorithms to predict customer churn. The script involved complex operations such as handling asynchronous API calls, processing JSON responses, and performing feature engineering on the data before feeding it into a machine learning model built with scikit-learn. The pipeline was automated to run on a schedule, with results stored in a database and visualized in a custom dashboard.  5. R  Example: In R, I worked on a time series forecasting project where I used the `forecast` package to develop a model for predicting future sales based on historical data. The data pre-processing involved handling missing values, applying transformations to stabilize the variance, and decomposing the time series into trend, seasonal, and irregular components. I also implemented cross-validation techniques to tune the model parameters and improve the accuracy of the forecasts, eventually deploying the model to provide actionable insights for the sales team. |
|  |