CSE 584

Summer 2023

40367

CPT Internship Report

BigCommerce

Amey Sadanand Bhilegaonkar

ASU ID: 1225368924

Master of Computer Science

Jessica Economou

Lead Data Solutions Developer

jessica.economou@bigcommerce.com

+1-512-839-0903

---------- 15/08/2023

Industry Mentor's Signature Date

---------- ----------

Faculty Chair's Signature (Thesis & Ph.D. students) Date

---------- 12/08/2023

Student's Signature Date

Student: By signing you certify that you have not plagiarized any of this document.

---------- ----------

Program Chair's Signature Date

Educational objectives of the internship:

This internship marked a significant milestone in my professional journey. My primary objective was to leverage my existing knowledge and delve deeper into the realm of Data Science. My focus was on engaging in hands-on experiences, encompassing the creation of data pipelines, proficiency in utilizing tools such as Python, Snowflake, Airflow, and SQL, and active participation within the agile development paradigm. My aspiration was to gain a comprehensive understanding of data management intricacies while contributing substantively to the team's ongoing initiatives.

Furthermore, a central objective of my endeavor was to contribute to the positive transformation within the company's operational framework. I endeavored to channel my enthusiasm and ignite creativity, whether by optimizing data pipelines for heightened efficiency, employing sophisticated SQL techniques for data refinement, or assuming an integral role in team deliberations. My overarching goal was to enhance the company's decision-making prowess through data-driven insights and introduce innovative strategies to optimize existing processes. At the same time, I aimed to build a mutually beneficial relationship with experienced professionals. This would help us share knowledge and skills, which would make me better at what I do and also help the organization do well.

Moreover, my immersive involvement in this internship enabled me to develop a multifaceted skill set that extended beyond technical proficiency. The collaborative nature of the projects allowed me to refine my teamwork and communication skills, as I actively contributed ideas, engaged in discussions, and aligned my efforts with the broader objectives of the team. The iterative nature of the agile development approach, with its daily standups and adaptive problem-solving, fostered an environment where I learned to be agile in my thinking and responsive to evolving project dynamics.

Simultaneously, my interactions with senior team members enriched my perspective and provided invaluable insights into industry best practices. The mentorship and guidance I received not only accelerated my learning curve but also underscored the significance of continuous learning and professional growth. Through candid discussions and knowledge-sharing sessions, I gained practical insights into translating theoretical concepts into actionable strategies and refining them within a real-world context.

Job Duties Included:

1. **Partnering closely with senior members of the Data Solutions team to design, and develop Machine Learning Model:** Collaborating hand in hand with esteemed senior members of the Data Solutions team marked a significant cornerstone of my enriching internship experience. This collaborative endeavor revolved around the strategic design and meticulous development of a sophisticated Machine Learning Model, solidifying my practical application of theoretical concepts. By working closely with seasoned professionals, I gained invaluable insights into the nuanced art of model planning, creation, and implementation. This comprehensive journey encompassed a meticulous understanding of project requirements, deliberation on the most suitable model architecture, rigorous data cleaning to ensure its integrity, and proficient coding of indispensable components. It culminated in the seamless deployment of the model for accurate outcome predictions. The symbiotic relationship forged during this process not only enriched my technical skills but also illuminated the significance of harmonious teamwork in achieving complex objectives.

2. **Creating and optimizing SQL transformations:** I was responsible for crafting SQL transformations that converted raw data into meaningful insights. This involved writing efficient SQL queries and optimizing them to improve performance, enabling faster data processing and analysis. I had to be clever and efficient with these instructions, so the computer could do its job quickly and give us the answers we needed.

3. **Participating in code reviews and performing deployments:** I actively engaged in code reviews, collaborating with my mentor to ensure high-quality code. Additionally, I created Google Colab notebooks, ensuring that changes were properly reviewed, tested, and integrated into the codebase to maintain a stable and reliable system.

4. **Working within an agile development environment participating in rituals and daily standups:** I embraced an agile development approach by engaging in various rituals such as daily standups. Through active participation in rituals such as daily standups, I fostered a heightened understanding of the symbiotic interplay between effective communication and cohesive teamwork. This agile framework galvanized me to navigate challenges with agility and adapt expeditiously to evolving project dynamics.

5. **Bringing fresh ideas to the table on automation and improvements:** I also brought some new and clever ideas to the table, like finding ways to make things work automatically and making our processes even better. I kept an eye out for ways to make things faster, more dependable, and full of creative thinking. I brainstormed ways to make things happen automatically and to make our processes even better than they were before.

   Company information:
   I worked for the company - BigCommerce. BigCommerce's mission is to help merchants sell more at every growth stage, from small startups to mid-market businesses, to large enterprises. As a leading Open SaaS solution, BigCommerce empowers merchants to build, innovate and grow their businesses online. Simply put, we focus on being the best commerce platform so our customers can focus on what matters most: growing their businesses.[1]

   BigCommerce was founded in 2009 and currently serves 150+ countries with tens of thousands of Merchants. BigCommerce is dedicated to its merchants, guided by our values and powered by their people. This is the driving force behind their success. It's what has taken them from a 2-person startup to a global team of hundreds helping retailers make billions of dollars in sales.

   The team I worked under was the Data Solutions team at BigCommerce. Data Solutions belongs to the Operations organization and acts as the technical bridge between the Analytics team and other cross-functional teams (Marketing, Legal, etc.). This team is responsible for designing and implementing data solutions that are technically efficient, repeatable, scalable, easy to maintain, and adaptable as BC's strategy expands.

   Data Solutions has a wide variety of responsibilities, including (but not limited to):
   - Integrating with our third-party tools to extract and model data (leveraging Airflow, Snowflake, and AWS)
   - Consulting on Analytics and Data Science projects, in addition to assisting with writing definitions of critical KPIs

- Iterating on our existing DAGs (directed acyclic graphs) to extract and transform data in an efficient manner. We also build new ones as new use cases arise!
- Tackle bug fixes on Snowflake / Airflow

Background about the problem
The project that I was engaged in encompassed the development of a comprehensive Machine Learning Classification model from its inception. The primary objective of this endeavor was to create a predictive model[7] capable of determining the likelihood of converting a customer who is currently on a trial period into a regular, paying customer. This predictive framework was aimed at providing valuable insights to guide strategic decision-making within the organization's customer retention strategies.

To embark on this ambitious undertaking, the initial phase involved rigorous data processing, where various raw datasets were meticulously curated and structured. This involved tasks such as data normalization, handling missing values, and addressing inconsistencies to ensure the dataset's quality and reliability. The comprehensive process of data cleaning was undertaken to guarantee that the subsequent analyses and model development were founded on accurate and reliable data.

Integral to the data preparation phase was the systematic removal of null values. Rigorous data cleansing methods were implemented to identify and subsequently eliminate instances of null values within the dataset. This process was crucial to enhancing the quality and integrity of the dataset, thereby ensuring the robustness of the eventual machine-learning model.

Moreover, the exploratory data analysis (EDA)[2] phase played a pivotal role in uncovering insightful trends and patterns within the dataset. Through a combination of statistical analyses, visualization techniques, and data mining, this phase aimed to gain a deep understanding of the relationships between various features and the target variable. Such insights provided a foundation for feature selection and engineering, thus informing the subsequent stages of model development.

In summation, this project entailed the holistic creation of a Machine Learning Classification model[4] with the ultimate goal of predicting the conversion potential of trial-period customers into regular patrons. The comprehensive data processing, cleaning, null removals, and exploratory analyses collectively served as the bedrock for developing an accurate and effective predictive model, facilitating strategic decision-making to optimize customer retention strategies.

Focus of the effort:
Approaching the intricacies of the project demanded a systematic and well-structured strategy. To tackle the challenge of building a Machine Learning Classification model for customer conversion prediction, I first embarked on a comprehensive literature review[6] to understand the underlying concepts and methodologies related to predictive modeling[7], data preprocessing[8], and exploratory data analysis. This step allowed me to gather valuable insights into best practices and emerging trends in the field, laying a solid foundation for my approach.

In order to effectively work on the problem, I had to enhance my proficiency in various domains. I delved into advanced concepts of data preprocessing, mastering techniques like data normalization, handling missing values, and outlier detection. Additionally, I strengthened my skills in exploratory data analysis, employing statistical tools and visualization libraries to uncover meaningful patterns within the dataset.

The knowledge acquired from previous coursework in machine learning, statistics, and data analysis proved to be invaluable, providing the theoretical framework necessary for problem-solving.

Addressing the problem demanded a versatile and thorough approach. It all began with a focused dive into the data preprocessing realm. I meticulously scrubbed and organized the dataset, leaving no stone unturned to ensure its reliability. The spotlight then shifted to feature engineering[9], a pivotal step where I conjured up new features from the existing ones, amping up the model's predictive prowess.

Taking a deeper plunge, I ventured into the realm of exploratory data analysis. Armed with visualizations and statistical tests, I went hunting for trends, links, and possible oddballs hiding within the features. This not only paved the way for smarter feature selection but also brought potential hurdles and breakthroughs to light, casting a revealing spotlight on the dataset's intricacies.

The core of the project lay in developing and fine-tuning[10] the Machine Learning Classification model. I experimented with various algorithms, iteratively adjusting hyperparameters and evaluating their performance using techniques like cross-validation. A major emphasis was placed on model interpretability, as understanding the factors influencing predictions was crucial for effective decision-making.

To measure the efficacy of the model, I employed a range of metrics[3] such as accuracy, precision, recall, and F1-score to gauge its performance on both training and testing data. Additionally, I conducted a comprehensive analysis of the confusion matrix to assess false positives and false negatives. This holistic evaluation allowed me to fine-tune the model further and optimize its predictive capabilities.

Incorporating feedback from rigorous code reviews and drawing inspiration from discussions with senior team members, I continuously refined the model, optimizing its accuracy and interpretability. This iterative refinement process not only elevated the model's predictive accuracy but also underscored the collaborative spirit and dedication to excellence within the project's ecosystem.

In conclusion, the successful approach to solving the problem hinged on a robust understanding of predictive modeling, data preprocessing, and exploratory data analysis. Leveraging the knowledge gained from coursework and supplementing it with extensive research, I applied a systematic strategy that encompassed data preparation, feature engineering, algorithm selection, and performance evaluation. The measurements made included a spectrum of performance metrics, enabling a comprehensive assessment of the model's accuracy and effectiveness in predicting customer conversions.

Results:
The culmination of my efforts yielded compelling results, showcasing an impressive accuracy rate of 97% coupled with high precision and recall values. This achievement underscored the robustness of the Machine Learning Classification model I had meticulously developed. The model's ability to predict customer conversion with such precision demonstrated its potential to significantly impact the organization's decision-making processes and customer retention strategies.

The efficacy of my solution resonated beyond development as it transitioned to practical implementation. The team recognized the value of the predictive model and subsequently proceeded to integrate it into the production environment. This seamless integration underscored the real-world utility of the model and its potential to influence operational strategies and customer engagement approaches.

While my efforts culminated in a successfully implemented solution, I did not write any formal papers documenting the project. The focus was primarily on the practical application and integration of the model, ensuring that its impact could be readily experienced and leveraged within the organization. The results and implications of the work were primarily communicated within the internal team, contributing to the ongoing refinement of data-driven decision-making processes.

Lessons Learned:

The internship proved to be an invaluable chapter in my academic journey, imparting a wealth of practical wisdom that reverberated through my academic experience. As I navigated the intricacies of the real-world data solutions realm, I found myself drawing extensively from the theoretical foundations laid in my academic courses. The strong understanding of core concepts gleaned from courses such as machine learning, statistics, and data analysis served as a robust scaffolding upon which I could build and expand my skills. The ability to seamlessly formulate and execute SQL queries, a skill honed through coursework, emerged as an indispensable tool during the internship's data processing and cleansing stages. This aptitude not only expedited data transformation but also amplified my confidence in handling complex datasets.

The internship was an immersive exercise in teamwork, solidifying my appreciation for collaborative dynamics in problem-solving. Collaborating closely with the Data Solutions team, I gleaned insights into effective communication, constructive critique, and shared decision-making. Working together on various stages of the project, from data preprocessing to model evaluation, highlighted the importance of harnessing collective expertise to tackle multifaceted challenges. This experience fundamentally redefined my approach to teamwork and underscored the vital role it plays in optimizing outcomes.

The most remarkable facet of the internship was the opportunity to construct a Machine Learning Classification model from the ground up. The hands-on nature of the project enabled me to holistically comprehend the intricate journey from data preprocessing to model deployment. This exercise was a testament to how academic knowledge could be translated into practical action. It empowered me to demystify the complexities of machine learning by systematically applying theoretical principles to real-world scenarios, resulting in an effective and accurate predictive model.

Furthermore, the internship equipped me with a gamut of skills that extended well beyond technical proficiencies. The ability to initiate and progress a project from scratch, while navigating and mitigating potential challenges, instilled in me a newfound sense of autonomy and problem-solving prowess. This experience honed my adaptability, enabling me to respond effectively to evolving project demands. Additionally, I honed my analytical thinking, leveraging insights from exploratory data analysis to shape informed decisions in the model development process.

In essence, the internship served as a conduit through which my academic knowledge gained tangible form. It reinforced the invaluable synergy between theoretical understanding and practical implementation, facilitating a seamless transition from classroom concepts to real-world applications. The exposure to collaborative dynamics, the orchestration of an end-to-end machine learning project, and the acquisition of multifaceted skills coalesced to foster holistic growth that transcended academic boundaries. The lessons learned and skills honed during this experience will undoubtedly shape my future pursuits and contribute to my overarching academic journey.

References:

1. BigCommerce Website: Link
2. EDA Medium Blog: Link
3. Accuracy, Recall, and Precision Blog: Link
4. My Data Mining Lab Assignment codes(Private Repo): Link
5. Logistic Regression Video tutorials: Link
6. Different Machine Learning Classification Models: Link
7. Predictive Modeling Tutorials: Link
8. Handling Null Values multiple GitHub repos: Link
9. Feature Engineering Blog: Link
10. Categorical Data Encoding Techniques: Link