राष्ट्रीय प्रौद्योगिकी संस्थान राउरकेला

NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

SHORT TERM INDUSTRIAL/RESEARCH EXPERIENCE REPORT

COMMMERCIAL VEHICLE CLAIM PREDICTION FOR

XXX COMPANY

Submitted by

XXXXXXXXXX
Bachelor of Technology
Computer Science & Engineering
National Institute of Technology
Rourkela, Odisha

Contents

Chapter 1

Abstract

I completed a two-month long internship at XXXgit General Insurance, where I worked as part of the Data Science and Analytics Team.

My main project was to apply the principles and practices of MLOps, which is a methodology that combines machine learning, DevOps, and data engineering, to improve the efficiency and quality of the company's data-driven solutions. In the first two weeks I learned about the domain of insurance and how to implement various machine learning models using MLOps tools and techniques.

Finally I developed an automated end-to-end pipeline for predicting commercial vehicle claims using various machine learning algorithms, MLflow, an open source platform for managing the machine learning lifecycle and FastAPI, a web framework for building APIs with Python. The pipeline enabled the company to deploy and monitor my models in a scalable and reproducible way, as well as to access the predictions and performance metrics through a user-friendly interface. My internship experience helped me gain valuable skills and insights in the field of Machine Learning,MLOps and its applications in the insurance industry.

Chapter 2

Introduction

The purpose of my internship was to gain practical experience in the field of Data Science and Analytics, and to apply the concepts and techniques of MLOps to real-world problems.

MLOps is a methodology that combines machine learning, DevOps and data engineering, to improve the efficiency and quality of data-driven solutions. It also aims to automate and streamline the entire machine learning lifecycle, from data collection and preparation, to model development and deployment, to monitoring and governance.

The organization I interned at was **XXX Company**, a leading digital insurance company in India. XXX offers various insurance products, such as health, motor, travel, and home insurance, through its online platform and network of partners. XXX leverages data and technology to provide innovative and customized solutions for its customers, such as personalized pricing, instant claims settlement, and smart device protection.

The context of my internship was to work on a project related to commercial vehicle insurance, which is one of the key segments of XXX's business. Commercial vehicle insurance covers the damages or losses caused by accidents involving vehicles used for business purposes, such as trucks, autos, taxis, etc.

The project involved developing an automated end-to-end pipeline for predicting commercial vehicle claims using machine learning models and MLOps tools. The pipeline would enable XXX to assess the risk profile of each vehicle, optimize the premium pricing, and expedite the claim processing.

2.1 Organisation Overview

XXX Company is a digital insurance company that was founded in 2016 by XYZ. The company's mission is to make insurance simple, transparent, and accessible for everyone. The company's values are customer-centricity, innovation, simplicity, and empowerment. The company's key activities are offering various insurance products, such as health, motor, travel, and home insurance, through its online platform and network of partners; leveraging data and technology to provide customized and convenient solutions for its customers; and simplifying the claim settlement process with features like smartphone-enabled self-inspection and zero paperwork.

The company has a strong data science vertical that is responsible for developing and deploying data-driven solutions for various business problems, such as risk assessment, premium pricing, fraud detection, customer segmentation, and claim prediction. The data science vertical consists of several teams that work on different domains and projects, such as health insurance, motor insurance, MLOps, etc. Each team has a manager who oversees the team members and coordinates with other stakeholders.

I worked under the manager of the data science team that was working on the MLOps project for commercial vehicle insurance. The manager's name was XXX Gupta, and he had over 10 years of experience in data science and analytics. He was very supportive and helpful throughout my internship, and he guided me in learning about the domain and the tools. He also gave me feedback and suggestions on how to improve my work and skills. He was a great mentor and leader for me.

2.2 Project Scope

The main scope of my internship was to work on the MLOps project for commercial vehicle insurance, which involved developing an automated end-to-end pipeline for predicting commercial vehicle claims using machine learning models and MLOps tools. The pipeline would enable XXX to assess the risk profile of each vehicle, optimize the premium pricing, and expedite the claim processing.

The pipeline consisted of four main components:

- **Data Ingestion and Preprocessing:** This component was responsible for preprocessing the data provided to make it clean and drop certain features to fit the model that has been put into production.

- **Model Training and Evaluation:** This component was responsible for building and testing various machine learning models, such as random forest classifier and catboost classifier to predict the probabilityt of claim for each vehicle. The models were trained and evaluated on different metrics, such as accuracy, precision, recall, F1-score. the idea behind this step was to automate the training whenever new data enters the system and put the best model based on some metric like F1 Score into production.

- **Model deployment and serving:** This component was responsible for deploying the best performing model to a production environment and serving the predictions through an API. The API was built using FastAPI, a web framework for building APIs with Python. The API allowed the users to input the vehicle details and get the predicted claim probability as output.

- **Model monitoring and governance:** This component was responsible for monitoring the performance and behavior of the model in production over time in comparison to other models stored in the MLFlow Registry.

2.3 Objectives

The objectives I aimed to achieve during the internship were:

- To learn about the domain of MLOps and its applications in the ever-growing field of machine learning.

- To implement various machine learning models using MLOps tools and techniques.

- To develop an automated end-to-end pipeline for predicting commercial vehicle claims using MLflow and FastAPI.

- To understand the insurnace industry and how things work inside an industrial setting.

Chapter 3

Internship Experience

I was working in the data science team at XXX Company, where I was involved in the MLOps project for commercial vehicle insurance. MLOps is a methodology that combines machine learning, DevOps, and data engineering, to improve the efficiency and quality of data-driven solutions.

3.1 Internship Description

My internship can be better explained by dividing it into multiple phases, the same being as follows:

3.1.1 Phase 01

In the first phase, I was asked to learn about the concepts of MLOps and about the values of the insurance industry. I studied various resources and tutorials on MLOps, I decided to focus on a tool in particular that is MLFlow. I practiced using these tools by creating and deploying simple machine learning models.

3.1.2 Phase 02

In the second phase, I was asked to use the MLOps techniques to make my own model for predicting commercial vehicle claims. The model had to be an automated end-to-end pipeline that would train and update itself with new data every month and put the best model into production. The pipeline also had to serve the predictions through an API and monitor the model performance and behavior over time. I used MLflow to manage the machine learning lifecycle and FastAPI to build the API.

3.2 Project Description

Commercial Vehicle Prediction is a project that aims to use various machine learning algorithms to forecast the probability of commercial vehicles being involved in insurance claims, aiding the company with risk assessment. The project can be described better using the following steps:

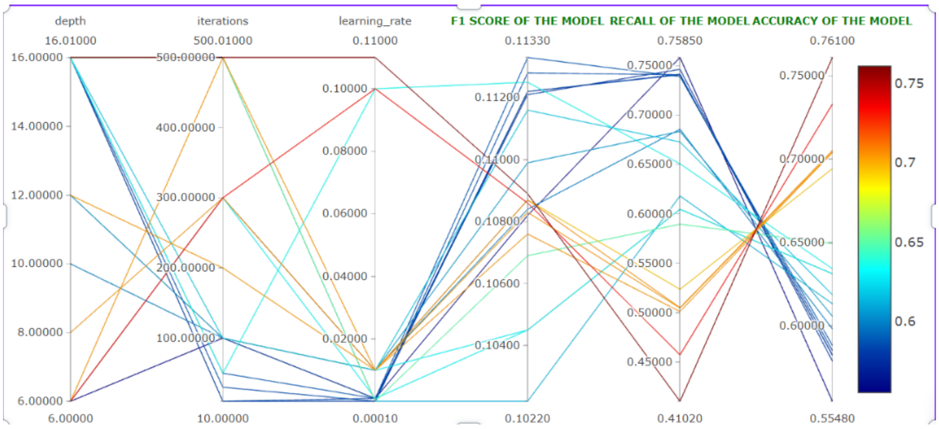
- **Data Collection:** The data for the project was obtained from the company's database, which contained 18 independent features and 1 dependent feature. The data included both numerical and categorical features, such as vehicle type, age, rto location, policy duration, claim amount, etc. The data spanned over a period of three years, from 2020 to 2023.

- **Data Preprocessing:** The data did not require a lot of cleaning, as it was already well-structured and validated. However, there were some missing values in two features: hypo party and prev insurer. Hypo party indicated whether the vehicle was financed by a third party, and prev insurer indicated the previous insurance provider of the vehicle. These missing values were imputed using a logical decision that NULL value for prev insurer suggested that XXXgit was the first insurance company it was going for and hypo party NULL value suggesting that the vehicle was self financed.
  A major task to handle was the imbalance of the dataset, as only 8 percent of the vehicles had claims. To deal with this issue, various resampling techniques were applied, such as random oversampling, random undersampling, SMOTE, etc., to create balanced datasets for training and testing.

- **Exploratory Data Analysis:** An in-depth analysis of the dataset was conducted to understand the distribution, correlation, and relationship of the features and the target variable. Various visualizations and statistics were used to explore the data and gain insights.

- **Model Selection:** Based on the literature review and domain knowledge, two machine learning algorithms were selected for the project: RandomForest and CatBoost classifier. RandomForest is an ensemble method that uses multiple decision trees to make predictions based on majority voting or averaging. CatBoost is a gradient boosting method that uses multiple weak learners to create a strong learner by minimizing a loss function. Both algorithms are suitable for handling categorical features and imbalanced data.

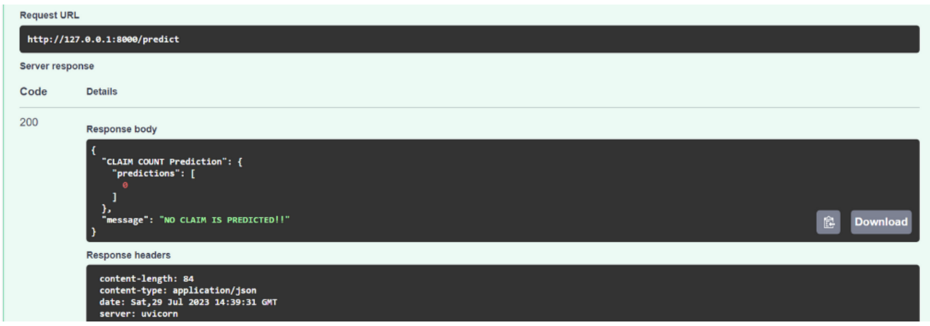
- **Model Training:** An automated training pipeline was created using MLflow, an open source platform for managing the machine learning lifecycle. The pipeline consisted of four stages: data loading, data splitting, model fitting, and model deployment. The pipeline took as input the raw data file and the model name, and outputted a trained model file and a MLflow run ID. The pipeline also logged various parameters and metrics for each model run, such as resampling technique, hyperparameters, accuracy, precision, recall, F1-score, etc., using MLflow tracking. The pipeline was run for both RandomForest and CatBoost models with different combinations of parameters and metrics.



- **Model Evaluation:** F1-score was chosen as the chief metric for evaluating the models, as it is a harmonic mean of precision and recall that balances both false positives and false negatives. The models were evaluated on both balanced and imbalanced test sets to compare their performance. The results showed that CatBoost had a higher F1-score than RandomForest on both test sets, indicating that it was more effective in predicting commercial vehicle claims.

- **CI/CD:** Continuous integration and deployment were implemented using MLflow as well. A scheduled job was set up to run the training pipeline every month with new data and update the best model based on F1-score. The best model was automatically deployed to a production environment using MLflow models and MLflow serving. The deployed model was exposed through an API endpoint that could be accessed by other applications or users.

- **Maintenance:** The deployed model was maintained using MLflow and FastAPI. MLflow was used to monitor the model performance and behavior over time using various metrics and alerts. FastAPI was used to create a web interface that allowed the users to input the vehicle details and get the predicted claim probability and amount as output. FastAPI also provided documentation and validation for the API endpoint.



Request URL

```
http://127.0.0.1:8000/predict
```

Server response

Code Details

200

Response body

```
{
  "CLAIM COUNT Prediction": {
    "predictions": [
      0
    ]
  },
  "message": "NO CLAIM IS PREDICTED!!"
}
```

Response headers

```
content-length: 84
content-type: application/json
date: Sat,29 Jul 2023 14:39:31 GMT
server: uvicorn
```

3.3 Learnings & Challenges

3.3.1 Business Perspective

- **Unveiling Data Complexity:**I encountered the reality of heterogeneous data in the insurance domain, which affects model performance. By slicing the data and understanding where the model works best, I fine-tuned my approach and achieved better predictions.

- **Challenges in Insurance Claim Predictions:**Predicting insurance claims is a complex task due to uncertainties and ever-changing risk dynamics.

- **Extracting Actionable Insights:**Visualizations of top features provided us with valuable insights into the factors influencing commercial vehicle claims.

3.3.2 Technical Perspective

- **Tackling Imbalanced Data:** By effectively addressing this issue, I strengthened my models' ability to handle positive outcomes and improved overall performance.

- **Embracing MLOps for Effciency:** Incorporating ML Ops practices significantly enhanced my development and deployment processes. I witnessed the benefits of reproducibility, scalability, and seamless collaboration.

3.4 Recommendations

Based on my experiences, I would like to offer some recommendations and suggestions for XXX Company to improve their data science and analytics practices and solutions. These are:

- To adopt a standardized and modular approach for MLOps.
- To follow the idea of treating interns as assets and not as a liability.
- To invest more in data quality and availability

Chapter 4

Conclusion

My internship at GoDigit General Insurance was a valuable and rewarding experience that helped me gain practical skills and insights in the field of data science and analytics, and MLOps in particular. I successfully completed the MLOps project for commercial vehicle insurance, which involved developing an automated end-to-end pipeline for predicting commercial vehicle claims using machine learning models and MLOps tools. The pipeline enabled the company to assess the risk profile of each vehicle, optimize the premium pricing, and expedite the claim processing. The pipeline also provided a user-friendly interface for accessing the predictions and performance metrics, as well as a robust mechanism for monitoring and governing the deployed model.

I am grateful to GoDigit General Insurance for giving me this opportunity to work on such an interesting and impactful project. I am also thankful to my manager Sujoy Sen Gupta for his guidance and support throughout my internship. I hope that my work will contribute to the company's growth and success in the insurance industry.

Acknowledgments

I would like to express my sincere gratitude and appreciation to the Career Development Center of National Institute of Technology Rourkela for providing me with this great opportunity to work as an intern at GoDigit General Insurance, a leading digital insurance company in India. This internship was a valuable and rewarding experience that helped me gain practical skills and insights in the field of data science and analytics.

I would also like to thank my manager Sujoy Sen Gupta, who guided and supported me throughout my internship. He was very helpful and generous in sharing his knowledge and expertise with me. He also gave me constructive feedback and suggestions on how to improve my work and skills. He was a great mentor and leader for me and the team.

I am grateful to GoDigit General Insurance for giving me this opportunity to work on such an interesting and impactful project.I hope that my work will contribute to the company's growth and success in the insurance industry.I learned a lot from this internship and I enjoyed every moment of it. I would highly recommend this internship program to anyone who is interested in pursuing a career in data science and analytics. I am confident that this internship will be beneficial for my future academic and professional endeavors.