AI-Powered Intelligent Insurance Risk Assessment and Customer Insights System

Overall Presentation

About Presentation:

1. What did I do in this project?

Built a **fraud detection model** for insurance claims using machine learning.

Developed a **fraud risk prediction app** with Streamlit.

Worked on **sentiment analysis using BERT embeddings** and **multi-language text summarization** using mBART/mT5.

Explored different ML and DL models, feature engineering, and hyperparameter tuning.

Experimented with **synthetic vs. real-world datasets** for fraud detection and NLP tasks

2. What did I learn?

Hyperparameter tuning improves model performance significantly.

Pretrained models (BERT, mBART) are better than custom embeddings for NLP.

Deep learning is not always superior to machine learning, especially for tabular data.

Real-world data is crucial for model accuracy—synthetic data often leads to poor predictions.

Outlier handling & feature scaling are essential preprocessing steps

3. What did I complete in this project?

✅ **Fraud Detection:**

- Built ML models (Random Forest, XGBoost) for fraud prediction.
- Deployed a fraud risk prediction app.
- Handled feature selection & scaling with MinMaxScaler.

✅ **Sentiment Analysis:**

- Tested ML models & deep learning (CNN+LSTM) on synthetic and real Twitter data.
- Achieved better accuracy using real-world data

✅ Text Summarization:

- Fine-tuned **mBART & mT5** for multilingual summarization.
- Switched to **BART-Small** due to hardware limitations.
- Improved ROUGE & BLEU scores using more real-time data.

4. What mistakes did I make?

❌ **Used synthetic data for fraud detection & NLP** → Poor accuracy due to randomness.

❌ **Over-reliance on deep learning for ML tasks** → Didn't improve accuracy significantly.

❌ **Initially ignored proper data preprocessing** → Later learned about outlier handling & feature scaling.

❌ **System crashes due to large transformer models** → Needed better resource management.

5. What did I learn from these mistakes?

✔ Synthetic data is unreliable for real-world predictions.

✔ Machine learning often works better for structured data than deep learning.

✔ Proper preprocessing (scaling, outlier handling) is crucial.

✔ Fine-tuning large models requires high computational power—choosing the right model matters

6.How did I rectify these mistakes?

🔄 **Switched from synthetic to real-world datasets** for better accuracy.

🔄 **Used ML models (Random Forest, XGBoost) instead of deep learning** for fraud detection.

🔄 **Implemented outlier handling (winsorization, trimming) and MinMax scaling.**

🔄 **Optimized transformer models** by using BART-Small instead of mBART/mT5.

7. Where did I face trouble in this project?

⚠️ **Fraud detection accuracy was poor** → Synthetic data was the problem.

⚠️ **Sentiment analysis models failed on synthetic data** → Lack of diverse words & context.

⚠️ **Training mBART/mT5 crashed the system** → Needed smaller models or better hardware.

8. How did I solve these troubles?

✅ **Used real insurance fraud & Twitter datasets** to improve accuracy.

✅ **Shifted to ML-based models for fraud detection** instead of deep learning.

✅ **Reduced transformer model size (mBART → BART-Small)** to avoid crashes.

9. How did I finish the project?

🎯 Finalized the fraud detection model with selected features & MinMax scaling.

🎯 Built a fraud risk prediction app in Streamlit.

🎯 Used real-time Twitter data for sentiment analysis.

🎯 Fine-tuned BART-Small for multilingual summarization.

🎯 Documented key learnings for future improvements

10. Overall thoughts on this project?

🚀 **Great learning experience!**

- **Improved understanding of fraud detection, NLP, and deep learning.**
- **Learned the importance of real-world data over synthetic data.**
- **Gained experience in hyperparameter tuning, feature engineering, and model selection.**
- **Realized that deep learning is not always the best choice for every task.**

11. My experience in Data Science

📌 **Strong foundation in ML & NLP models**

📌 **Worked on fraud detection, sentiment analysis, and text summarization**

📌 **Experience with ML models (Random Forest, XGBoost) & deep learning models (BERT, mBART, LSTMs)**

📌 **Hands-on with data preprocessing, feature selection, and hyperparameter tuning**

📌 **Building & deploying real-world ML applications (Streamlit app for fraud detection)**

12. What topics am I strong in currently?

✅ Machine Learning models & evaluation metrics

✅ BERT embeddings & NLP model fine-tuning

✅ Preprocessing techniques (feature scaling, outlier handling, text tokenization)

✅ Fraud detection and sentiment analysis

13. What topics do I need to concentrate on deeply and why?

🔍 **Transformer models & advanced NLP** → Need better fine-tuning skills for large models like mBART.

🔍 **Optimization techniques for deep learning** → Improve loss reduction strategies.

🔍 **Feature engineering for structured data** → Better fraud detection models with high-impact features.

🔍 **MLOps & model deployment** → Learn how to deploy models efficiently in production.

Translation vs Summarization

Feature	Machine Translation (MT)	Text Summarization
Input Format	`translate English to French: {text}`	`summarize: {text}`
Output Length	Similar length as input	Much shorter than input
Evaluation Metric	**BLEU score** (fluency, accuracy)	**ROUGE score** (content overlap)
Training Data	Sentence pairs from parallel corpora	Long-form text & summaries
Challenges	Maintaining meaning & grammar	Compressing while keeping key info

Conclusion:

This project was a great journey in **fraud detection, NLP, and deep learning**. I learned the **importance of real-world data, proper model selection, and optimization techniques**. Moving forward, I will **focus on advanced NLP, transformer optimization, and MLOps** to take my skills to the next level! 🚀

Thank You