HARNESSING AI AND ML IN OFFICIAL STATISTICS:

THE SORS EXPERIENCE

Transforming Statistical Operations through Advanced Technology at the Statistical Office of the Republic of Serbia (SORS)

Marko Grujičić
Adil Kolaković

Introduction to AI and ML in Official Statistics

- AI and ML are revolutionizing official statistics by automating complex processes, improving data accuracy, and enabling real-time analysis.

- These technologies have advanced from basic machine learning models to sophisticated systems like Large Language Models (LLMs).

- The integration of AI/ML not only enhances efficiency but also opens new possibilities for using unconventional data sources, contributing to timely statistical outputs.

Overview of SORS's Digital Census

- The 2022 Census marked Serbia's first fully digital census, integrating AI and ML throughout the data collection and processing stages.

- SORS enhanced data accuracy, reduced costs, and sped up the census process by adopting a digital approach.

- This digital transition set a new standard for census operations in Serbia, showcasing the potential of AI and ML in revolutionizing statistical practices.

ML in the Post-Enumeration Phase

- SORS used Machine Learning to classify occupations and economic activities based on ISCO and NACE classifications.

- ML automated the traditionally labor-intensive classification process, ensuring quicker and more accurate results.

- The models were trained on large datasets that included variables like education, age, and gender, improving classification precision.

Data Anonymization and Cloud Processing

- SORS used cloud-based infrastructure to efficiently process large datasets for ML tasks, ensuring scalability and security.

- Data was anonymized before cloud processing to protect sensitive information, maintaining respondent privacy.

- The cloud environment, equipped with high-performance hardware, allowed SORS to handle complex ML algorithms swiftly.

ML Model Training and Data Sources

- Training began with data from the 2011 census, enhanced with ongoing surveys like the Labour Force Survey and the Central Register of Mandatory Social Insurance (administrative source in Serbia).

- These diverse data sources provided a robust training set for accurate classification of occupations and activities.

- Combining multiple data sources allowed SORS to create a reliable ML model, addressing potential biases with varied demographic factors.

IST’s Role in Enhancing AI/ML Efficiency

- Seamless Data Collection:

IST's advanced functionalities, including efficient data collection and robust logical controls, ensured that high-quality, clean data was available for AI/ML processing.

- Modern Architecture:

IST's architecture, built on MS SQL DB Server, provided a stable and scalable foundation, minimizing the need for extensive preprocessing before AI/ML applications.

- Improved AI/ML Performance:

By starting with well-structured, accurate data, IST significantly enhanced AI/ML performance, reducing resource requirements and ensuring project viability.

- Avoiding GIGO:

IST’s capabilities helped us avoid the common pitfall of 'Garbage In, Garbage Out,' leading to more reliable and actionable insights from our AI/ML models

ML Algorithms Used

- SORS tested and selected the Random Forest classifier for its superior accuracy in classifying occupations and activities.
- The Random Forest model achieved 98% accuracy, a critical factor in ensuring high-quality census data.
- This algorithm was ideal for handling the diverse and complex data collected during the census.

Achievements in Classification Accuracy

- The Random Forest classifier improved occupation and activity classification accuracy, achieving a 98% accuracy rate.
- This high accuracy was validated against manually coded samples, ensuring reliable results.
- These improvements enhanced the overall quality of the census data, providing a solid foundation for future statistical operations.

Practical Benefits of AI and ML in the Census

- AI and ML reduced processing time and errors by automating classification tasks, ensuring consistent data quality.

- These technologies enabled SORS to handle large datasets more efficiently, making the census process faster and cost-effective.

- The integration of AI and ML in the census highlighted their potential for broader applications in official statistics.

Challenges Faced and Lessons Learned

- Challenges included data quality issues with OCR text and the computational demands of ML model training.

- SORS developed secondary ML models for text correction and optimized cloud resources to address these challenges.

- These solutions provided valuable lessons for future projects, equipping SORS with the knowledge to enhance AI and ML use.

AI and ML Applications in Global NSOs: Statistics Canada

- Statistics Canada uses ML to automate survey response coding, resulting in faster processing times and improved data accuracy.

- ML reduced the need for manual coding, minimizing errors and streamlining operations.

- This approach demonstrates the versatility and effectiveness of ML in the statistical domain.

AI and ML Applications in Global NSOs: ONS (UK)

- ONS in the UK has used AI for business activity classification, improving the accuracy and timeliness of economic statistics.

- AI applications enabled ONS to process and analyze large datasets more efficiently, providing reliable data for economic policies.

- This success highlights AI's potential to revolutionize statistical data collection and processing.

Future Directions for AI and ML at SORS

- SORS plans to expand AI/ML use to other statistical domains, including environmental and social statistics.

- Future initiatives will involve collaboration with other NSOs and international organizations to share best practices.

- SORS is committed to leading AI and ML innovation in statistics, enhancing data accuracy, efficiency, and relevance.

Ethical Considerations & Best Practices

- Adherence to global ethical standards ensures responsible and transparent use of AI/ML in official statistics.

- SORS has developed guidelines for ethical AI deployment, focusing on fairness, accountability, and transparency.

- These guidelines prevent biases, protect privacy, and ensure AI benefits to all stakeholders in the statistical process.

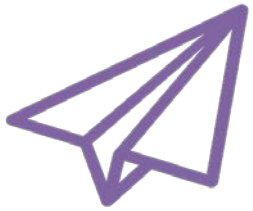
Conclusion and Q&A

- SORS's 2022 census demonstrated the transformative power of AI/ML in official statistics, setting a new standard for digital data collection.

- SORS is committed to continuing AI/ML innovation, working with global partners to advance these technologies in statistics.

- This presentation highlighted SORS's achievements and challenges with AI/ML, and we welcome any questions or discussions.

Contact Information



Email: ist@stat.gov.rs

Website: https://istportal.net/