REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS

ANNEX I. Technical Specification Double-side Page

TECHNICAL SCOPE

The project focuses on creating a robust decision-making system designed to optimize online marketing campaign performance on Google.com and its related networks. By analyzing various aspects of campaign configurations and keywords, the system generates data-driven recommendations for the most effective networks and keywords to launch new campaigns. This solution is specifically tailored for startups seeking to establish and enhance their online marketing strategies. Utilizing advanced machine learning and deep learning techniques, the system delivers a comprehensive and intelligent approach to campaign optimization.

Modalities:

- Express diagnosis: Requires minimal campaign information for diagnosis and optimization, with outputs being optimization recommendations and pre/post optimization performance predictions.

Input: Category, Country, Keyword and Network. Note: Language, Match type, Currency, and Device are temporarily discarded while gathering more information.

Output: Campaign performance predictions and campaign suggestions optimised by ML models maximising CTR.

- Tailored engine: Requires uploading user/startup information and previous campaign data (variables to be studied in Experiment phase), resulting in better recommendations and performance through transfer learning.

Solution Workflow:

- Scenario Exploration: Modify the proposed campaign by testing various categories, networks, countries, and keywords. New keywords are generated using a generative large language model and the best matches are selected.

- Scenario Evaluation: Proposed scenarios are evaluated using a machine learning model.

- Suggestions: Top-performing campaigns are presented to the user as well as the performance prediction (CTR)

SELECTION OF ALGORITHMS AND TOOLS

1. Model Development Tool: Python, due to its extensive machine learning libraries, ease of use, and widespread adoption.

2. Implementation Tool: Google Cloud Platform: GCP offers a robust infrastructure, scalability, and integration with various data services, making it the ideal choice for model training, deployment, and management.

3. Development Tool: Google Colab: Colab provides a browser-based Python environment that allows easy collaboration, version control, and access to powerful GPUs, along with seamless integration with Google services.

4. MVP/Mockup Tool: Google Colab: Offers a user-friendly interface for quick feedback and iteration during the prototype stage.

5. Frontend: API/Website: The final product will include an API and website for easy integration and a smooth and simple user experience.

6. Machine Learning Models and Algorithms:

- **Linear Regression**: Linear regression is used to gain insights from the feature weights as a starting point. It was chosen for its simplicity and interpretability.

- Fully Connected Neural Network: A fully connected NN is used to predict the campaign performance given a specific campaign configuration. It was chosen for its ability to model complex relationships and non-linear patterns between input features and CTR, capturing intricate patterns within the data.

Due to the availablility of huge amounts of data in the next phases, new targets will be included to make the neural net multitask which also will improve it's performance letting this solution be part of a multistakeholder Data value chain.

- **Word embeddings**: Word vector embedding is used for feature engineering tasks, encoding keywords and categories into vector embeddings. It was chosen for its effectiveness in capturing semantic relationships between words, which allows for improved keyword similarity calculations and assists in generating relevant keyword suggestions for campaign optimization.

- **Generative Pretrained Transformer**: GPT is used for feature engineering tasks, generating new keywords based on the current keywords and the campaign context. It was chosen for its ability to create diverse, contextually relevant keyword suggestions, helping users explore alternative keywords that may improve their campaign performance.

- **Cosine Similarity**: Cosine similarity is used for feature engineering tasks, measuring the relationship between categories and keywords, and selecting the best keywords for the campaigns. It was chosen for its effectiveness in comparing the similarity between vectors, such as the Word2Vec embeddings, enabling a data-driven approach to keyword and category selection.

TECHNICAL SCALABILITY AND FLEXIBILITY OF THE SOLUTION

Scalability:

The solution's scalability comes from its cloud-based infrastructure, distributed data processing, advanced machine learning libraries, and API-driven design, ensuring efficient handling of large datasets and seamless adaptability across various domains and data providers.

Cloud-Based Infrastructure: To handle large datasets and scale compute resources, the solution is hosted on Google Cloud.

Distributed Data Processing: The solution will use Apache Spark to process and analyze large datasets across multiple nodes in parallel, thus efficiently handling humongous data.

Advanced Analytics and Machine Learning: Scalable machine learning libraries were used like TensorFlow. This allows the solution to perform advanced analytics on large datasets and adapt to different domains by fine-tuning and retraining models as required.

API-driven Design: An API-driven design is implemented, that makes it easier to integrate with various data providers, adopt new data sources, and support integration with other platforms and services in different domains.

Flexibility:

The solution ensures flexibility through adaptive machine learning models, a robust data pre-processing pipeline, transfer learning techniques, open-source frameworks, and cross-platform compatibility. This enables seamless integration with diverse data providers and adaptation to various marketing and advertising use cases.

Adaptive machine learning models: The selected machine learning models, such as linear regression and fully connected neural networks, can be easily retrained with new data, ensuring that the performance of the models remains accurate and up-to-date with changing market dynamics.

Robust data pre-processing pipeline: The data preprocessing pipeline is designed to automatically handle missing values, inconsistencies, and outliers, enabling the seamless integration of datasets from diverse data providers.

Transfer learning: The solution leverages transfer learning techniques to use pre-trained models or features for different tasks or domains. This allows for faster model training and adaptation to other related use cases in the marketing and advertising domain.

Open-source frameworks and libraries: The solution leverages popular open-source frameworks and libraries, such as TensorFlow and scikit-learn, which facilitates easy integration with other technology stacks and contributions from the data science community.

Cross-platform compatibility: The solution is designed to be platform-agnostic, ensuring its compatibility with different systems and devices for maximum flexibility and reach to users.

DATA GOVERNANCE AND LEGAL COMPLIANCE: Data sharing challenges, data governance and legal compliance, must be observed. The proposed solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

Our solution emphasises Data Governance and Legal Compliance by utilising Google Cloud Platform's secure infrastructure. We will manage encryption, access control, and user consent management to address data sharing challenges and maintain compliance with GDPR. Below, we outline key technical aspects that demonstrate our compliance.

Data Storage and Encryption: All sensitive data is securely stored in the Google Cloud Platform (GCP) storage services, such as Google Cloud Storage or Google Cloud SQL. Data encryption is applied both at rest and in transit, utilizing GCP's built-in encryption methods like server-side encryption with Google-managed encryption keys or customer-managed encryption keys.

Access Control and Identity Management: Robust Identity and Access Management (IAM) policies are implemented to manage and restrict access to data based on user roles and privileges. This ensures that only authorized personnel can access and modify the data, minimizing the risk of unauthorized data access or manipulation.

User Consent and Data Usage Control: To ensure transparency and that users maintain control over their data, the solution provides users with an opt-in or opt-out mechanism for the usage of their data in training new machine learning models. This is achieved through a clear and intelligible consent form that outlines the purpose, benefits, and potential risks associated with using their data for model training.

QUALITY ASSURANCE AND RISK MANAGEMENT

Our approach focuses on addressing potential risks and maintaining high quality throughout the design, development, testing, and deployment phases of our online marketing campaign optimization solution. By incorporating rigorous testing, performance monitoring, and continuous improvement measures, we ensure a robust and reliable product for small businesses and startups.

1. Design Phase

- Conducted research to identify challenges and set project goals.

Risk Mitigation: Engaged stakeholders for feedback.

2. Development Phase

- Implemented data preprocessing and feature engineering techniques, trained machine learning models.

Risk Mitigation: Validated models using cross-validation and performance metrics.

3. Testing Phase

- Conducted extensive testing and ensured compliance with data governance, privacy, and security regulations.

Risk Mitigation: Updated development and testing plans based on identified risks.

4. Deployment and implementation Phase

- Deployed the solution on Google Cloud Platform and developed a user-friendly frontend.

Risk Mitigation: Implemented performance monitoring and logging mechanisms.

5. Continuous Improvement

- Establish ongoing quality assurance and risk management frameworks.

Risk Mitigation: Conducted periodic audits and reviews to maintain compliance and minimize risks.

REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS