ANDREW CHAUZOV

Data Scientist, Machine Learning Engineer, Data Analyst

Location: France

avchauzov@gmail.com /avchauzov /avchauzov

- Boasting over 10 years of experience in Data Science & Analytics, with 7 years in Python & Machine Learning and 5 years in Deep Learning & Engineering.
  Demonstrated proficiency in analyzing and modeling datasets with millions of rows, developing over 30 complex machine learning models in recent years.

- Recently focused on NLP, showcasing skills in BERT, Transformers, Transfer Learning, and LLM APIs.
  Employed these techniques to categorize over 250k social media accounts, improving customer engagement strategies.

- Strong foundation in machine learning methodologies and statistics.
  Implemented hypothesis testing methods to refine client targeting, achieving a 3x fine-tuning of target groups.

- SQL, large datasets management, API development, and deploying solutions on GCP/AWS.
  Enhanced the speed of a predictive model, markedly improving user experience.

- Proficient in communicating complex data insights and mentoring.
  Led and mentored a team of 3, driving solutions that increased client revenue by 10%.

Experience

03/2020 – Present: Freelance Artificial Intelligence Engineer at **TOPTAL** - Digital platform connecting skilled talent with global business opportunities.

- Formulated a **BERT**-based approach for Instagram profile categorization, achieving over **80% accuracy** across **50+ groups**, enhancing ad campaigns.
- **Engineered** over **15 predictive models** for **100+ million IDs** each via **GCP** using **PL/SQL** and **regex**, ensuring **scalability** and efficient **logging**.
- Implemented audience **enrichment** with **hypothesis testing** and **SQL**, generating over **100k** potential customer matches per company.
- Employed a **nearest neighbors** model for Twitter account category estimation (~4*16 subclasses) using **TFHub BERT embeddings**.
- Leveraged **ChatGPT API** for **data generation**, standardizing job titles and estimating seniority/department with **~90% accuracy**.
- Applied **random walk embeddings**, **UMAP**, and **HDBSCAN** to customer interests, enriching a targeted ad subset for **25k users**.
- Designed a **Scikit-Learn** salary prediction pipeline with less than **10% error** using **unsupervised transformation**.

11/2020 – 10/2023: Data Scientist at **CRED INVESTMENTS** - Platform for investing in and nurturing emerging talent.

- Combined **classification analysis** and **tree embeddings** for in-game location data, identifying player roles with **~90% accuracy** and improving **clustering**.
- Assembled a 'team profile' algorithm using **gradient boosting** and **statistics**, achieving **85% accuracy** in pinpointing team weaknesses within a league.
- Collaborated on a **25-page Streamlit dashboard** featuring **Plotly**, **Seaborn**, and **Matplotlib** visuals; actively participated in **GitHub** code reviews.
- Deployed over **10 predictive models** via **REST API**, managing over **100k predictions** with rapid response times and robust data issue resolution.
- Designed a dual-layer **regression model** for football market value forecasts, achieving short and long-term predictions with **<10% error**.
- Created a 4-step **data engineering** pipeline for player stats (**cleaning**, **scaling**, **imputation**), enhancing models' accuracies by **~75%.**
- Architected a **LightGBM model** with a **15% error rate**, identifying promising young talents and **suggesting optimal replacements**.

01/2017 – 03/2020: Data Scientist at **FREELANCE**

- Applied **NumPy** to custom sale volumes **clustering algorithm**, analyzing over **10k time series**, resulting in up to a **15% increase** in sales effectiveness.
- Trained a **CatBoost** model for employee **churn prediction** task with **85% accuracy**, providing HR with actionable insights using the **LIME** library.
- Achieved **70-95% accuracy** in detecting speech defects in children; deployed the model for a mobile app using **FastAPI** and **Docker**.

02/2012 – 12/2016: (Senior) Data Analyst at **ASSOCIATION 'NON-PROFIT MARKET COUNCIL'** - Central hub fostering collaboration and trade in energy markets.

- Conducted **ARIMA-based time series analysis** for **anomaly detection**, leading to a **50% increase** in prediction accuracy and improved data quality.
- Enhanced power price/volume **forecasting models** using **feature engineering**, achieving up to a **2.5x reduction** in error rates for targeted regions.
- Led **4 commercial predictive projects** with **teams of 2-4**, successfully deploying models that **enhanced client revenues** by approximately **10%**.

07/2011 – 02/2012: Junior Data Analyst (Marketing Mix Modeling) at **BBDO GROUP** - Top-tier network for creative marketing and advertising.

- Implemented an **anomaly detection function**, leading to a **50% reduction** in **expenditure** and **improved** client **marketing budget efficiency**.
- Introduced **automated scripts**, **increasing** departmental **task efficiency 4x** through streamlined media **data collection** and **processing**.
- Initiated **cluster analysis** to estimate pre-campaign efficiency, leading to more effective budget allocation across over **10 campaigns**.

Projects

08/2020 – 07/2022: ML Development & Engineering for Hong Kong Horse Race Prediction Algorithm

- Developed efficient data processing scripts and enhanced **feature engineering** with **TsFresh**, adding over **25 critical features** to the dataset.
- Conducted **classification analysis** with accuracies ranging from **65% to 90%** for different outcomes, optimizing bet amounts with **Hyperopt**.
- Orchestrated an **AWS**-based pipeline with **MySQL** for data storage, delivering real-time predictions **every 30 seconds** via **Flask**.

Education

Peoples' Friendship University of Russia

Master of Science - MS - Applied Mathematics & Computer Science

Skills

- Python (Pandas, SciPy, NLTK, spaCy), PyTorch, TensorFlow, Keras, Jupyter Notebook, Git, Frameworks (MLflow, Neptune.ai), Luigi, Gradient Boosting (XGBoost, LightGBM), NLP (Text Analysis, Text Processing), Optimization Techniques (Optuna), Data Drift.
- Exploratory Data Analysis (EDA), Data Cleaning, Data Mining, Data Manipulation, Data Validation, Data Analytics, Data Modeling, Data Visualization, Predictive Analytics, Statistical Modeling.
- Bayesian Statistics, Probability Theory, Statistical Analysis, Dimensionality Reduction (PCA), SHAP, Permutation Importance.
- Supervised/Unsupervised Learning, Network Analysis, Graph Theory.
- Databases (BigQuery, Redis), Cloud Computing (Amazon Web Services, Google Cloud Platform).
- Linux (Ubuntu, Debian), Stakeholders Interaction, Reporting, Scraping (Selenium).
- Model Development, Model Selection, Model Validation, Model Deployment, Software Development Practices, Agile (Scrum).