Kerem Burak Yılmaz

Istanbul kyilmaz22@ku.edu.tr 0 531 379 28 91 Kerem Burak Yılmaz keremburakyilmaz

Summary

- Passionate Computer Engineering student (GPA: 3.86) with hands-on experience in machine learning, MLOps, generative AI and quantitative analysis.
- Designed and deployed full-stack AI systems including a Retrieval-Augmented Generation (RAG) chatbot for business solution discovery and end-to-end ML pipelines for financial forecasting, customer churn prediction, and F1 race classification.
- Skilled in deploying applications with Docker, FastAPI, and AWS EC2/RDS, and integrating model monitoring with MLflow and feedback-aware fine-tuning loops.
- Strong foundations in object-oriented programming, data structures, UI/UX, and teamwork through diverse collaborative projects and hackathons.

Education

Koç University *BEng in Computer Engineering, GPA: 3.86* *Sept 2022 – June 2026*
Vehbi Koç Honor List

Koç University *BBA in Business Administration, GPA: 3.86* *Jan 2024 – June 2027*
Vehbi Koç Honor List

Experience

Co-founder *Remote*
AINA *Oct 2024 – Present*

- Developing our hackathon-winner idea into a business.
- Building a mobile app for entertainment of clothing customers where they can use AI to rate their outfits and get recommendations, create outfits with their clothes, and find shopping options to add to their closet.
- Responsible for the full-stack Flutter + Supabase, the recommendation model and the rating system.

AI Engineering Intern *Remote*
Exin Health AI *June 2025 - Present*

- Developing an IOS mobile app that digitalizes operation rooms and makes it easier to fill out forms using ASR and LLMs.
- Working on how to decrease the hallucination on LLM outputs and how to get more accurate JSON outputs based on the recognized speech input.

AI Engineering Intern *Hybrid - Beyoglu/Istanbul*
Digitopia *May 2025 - Present*

- Working on a chatbot to help customers understand which DMI level they are currently at and what steps they should take in order to achieve their goals.

AI Engineering Intern *Remote*
Genarion *Feb 2025 – May 2025*

- Worked on software applications based on LLM.
- Wrote an interview script which generates questions evaluating multiple areas including hard and soft skills based on the given job post, CV, and previous answers using TTS and STT.
- Finetuned multiple TTS models to speak Turkish naturally with self written scripts to create dataset from youtube videos automatically.

Machine Learning Intern *Remote*
Forma Makine *Dec 2024 – Feb 2025*

- Learnt about statistics and mathematics of the machine learning algorithms.

Projects

Business Solution Discovery Chatbot (RAG-based)

- Designed an end-to-end chatbot using Retrieval-Augmented Generation (RAG) to match users with tailored business solutions.
- Implemented semantic search with FAISS and SentenceTransformers; integrated LLaMA-3 via Groq API for real-time generation.
- Built a feedback-aware retraining pipeline using MySQL logs and CosineSimilarityLoss.
- Automated web scraping with Selenium and developed analytics dashboards with Matplotlib and Pandas.
- Deployed production backend on AWS EC2 and managed cloud database with RDS.

QuantFusion – AI-Powered Financial Intelligence Platform

- Building a modular AI-powered finance platform integrating portfolio optimization, risk analytics, sentiment analysis, and algorithmic trading into a unified backend–frontend system.
- Developed FastAPI-based risk analysis supporting VaR, CVaR, volatility, drawdown, CAPM beta, and risk attribution using both historical and parametric methods.
- Implemented advanced portfolio construction strategies including Mean-Variance Optimization (Markowitz), Risk Parity, and Black-Litterman with real-world constraints like sector limits, weight bounds, and tracking error.
- Designing a React frontend for real-time dashboards, visualizations, and interactive model control.
- Roadmap includes deployment of options pricing models, a sentiment-driven market forecasting engine, and a rule-based trading module.

ChurnSight – End-to-End MLOps Pipeline

- Developed a complete end-to-end MLOps pipeline for customer churn prediction with custom implementations of Logistic Regression, Decision Tree, Random Forest, XGBoost, MLP, and Gaussian Naive Bayes classifiers, and implemented a meta classifier using all the custom models.
- Automated hyperparameter tuning with Optuna and evaluated models using ROC-AUC and accuracy.
- Deployed FastAPI inference API with support for batch predictions and SHAP-based feature explanations.
- Containerized the app with Docker and integrated basic CI/CD workflows via GitHub Actions.

F1 Predictor – Driver Outcome Classification

- Built a multithreaded data pipeline with FastF1 and ThreadPoolExecutor to collect historical F1 race, weather, and qualifying data.
- Trained a custom Random Forest classifier to predict driver categories (Top 3, Midfield, Backmarker) with 77% accuracy.
- Achieved 100% recall on Top 3 predictions; validated model with precision/recall scores and confusion matrix analysis.

Languages

Turkish: Native
English: C1 (Advanced) - KUEPE: 91