Artiom

Data Scientist, BI Engineer, Data Engineer

Key strengths:

- ML and data science programming using Python, Scala, Golang, R.
- Data analysis and machine learning skills. DWH, ETL processes.
- Deep learning skills.
- Knowledge of Telecom, E-commerce, Banking & Finance and Retail domains. Requirements gathering and documenting skills

7+ years of experience

EDUCATION

University	Belarusian State Economic University
Department	Applied mathematics
Degree	Master

MAIN PROFESSIONAL SKILLS

Programming Languages	Python Golang Matlab R SQL Scala
	Java
Databases, cloud platforms	Mongo BigTable HBase Cassandra Oracle RDBMS PostgreSQL MySQL MariaDB MS SQL Server ClickHouse Vertica Elasticsearch Google BigQuery Google Cloud Platform AWS EC2
Tools	SPSS Eviews Microstrategy BI Oracle BI Tableau Git Jenkins SVN

Technologies	Apache Spark Flask Pandas Numpy Scipy Scikit-learn Keras PyTorch TensorFlow

LANGUAGES

Language	Listening	Speaking	Reading	Writing
English	Upper-Intermediate	Upper-Intermediate	Upper-Intermediate	Upper-Intermediate
French	Basic knowledge	Basic knowledge	Basic knowledge	Basic knowledge

RECENT PROJECTS

Project #1

Name of the project	**Real-time invoice processing data pipeline**
Description	Real-time data pipeline built in AWS cloud (step functions + lambda) to parse and process invoice PDF files. Invoice insights (analytics) implemented using Apache Spark. Postgres was used as the main database. The backend API was built using Python FastAPI.
Role	Senior Data Engineer/Backend Engineer
Tasks executed	• OCR/NLP - correct invoice parsing • Canonical mapping of invoice vendors/line items • Data pipeline with reprocessing and validation functionality and logging to Elasticsearch
Technologies used	AWS cloud (step functions, lambda-serverless, ECS, Fargate), Docker, Spark, Postgres, Elastic

Project #2

Name of the project	**Recommender system for OSS components based on GitHub data**
Description	One of the main goals of the product was to implement a recommendation system which suggests a potential user to follow, compare or analyze similar components with high quality to those already chosen previously. Recommendations were based on both structural similarities between components (NLP of components readmes, title, files & folders structure, then keywords extraction) and usage analytics with defined patterns of behavior on the frontend extracted from yandex metrics raw data and preprocessed in python. ML algorithms and data science stuff were developed on python. Data storage and collection were developed in Mongo for totals and github crawled data and Bigtable for time series.
Role	Chief Data Scientist/Business Analyst
Tasks executed	● Accurate and reliable NLP of readmes and title ● Specific keywords extraction without general and commonly used words ● Detection of patterns of behavior in usage analytics raw data, segmentation of users based on these patterns
Technologies used	Python, Mongo, Bigtable, BigQuery, Docker, GitHub

Project #3

Name of the project	**Keywords generation for OSS components based on github data**
Description	One of the key goals of the product was development of categories/tags for search engine to make the search of components for users more user-friendly and easier. It was decided to extract these set of tags from the title and readme file of github components. As soon as they were crawled and stored in Mongo, they were processed using several packages like mistune, beautiful soup, nltk and textacy to get only specific tags related to the component without commonly used words and junk words. These keywords were then used in search engine. ML algorithms and data science stuff were developed on python. Data storage and collection were developed in Mongo for totals and github crawled data and Bigtable for time series.
Role	Chief Data Scientist/Business Analyst
Tasks executed	● Accurate and reliable NLP of readmes and title ● Specific keywords extraction without general and commonly used words
Technologies used	Python, PyTorch Mongo, BigTable, BigQuery, Docker, GitHub

Project #4

Name of the project	**Road crash prediction model**
Description	Classification model to predict an incident. Roads were aggregated to clusters, each cluster had a bunch of metrics, like road sensors, weather information, road characteristics, road constructions, etc. Severe class balancing problem was solved using an imblearn combination of under- and oversampling methods.
Role	Chief Data Scientist
Tasks executed	● Solved class balancing problem ● Low level of false negatives in prediction
Technologies used	Python, Mongo, Bitbucket

Project #5

Name of the project	**Real estate assets price prediction**
Description	Prediction of next real estate price using regression model approach. The model was implemented using scikit-learn pipelines.
Role	Chief Data Scientist
Technologies used	Python, PostgreSQL, GitHub

Project #6

Name of the project	**Prediction of sell intent for real estate assets**
Description	Binary classification task regarding prediction for a specific property to be sold or not. Data available: 70K properties with historical information about loans, previous owners, other statistics.
Role	Chief Data Scientist
Technologies used	Python, PostgreSQL, GitHub

Project #7

Name of the project	**Categorization of OSS components based on GitHub data**

Description	Github currently doesn't provide a categories feature for components. The task was to label the initial dataset using a set of predefined categories, depending on the content of the repository and then build ML model, predicting the specific class of the repository based on text data available: repository description, readme file, files, and folders hierarchical structure of the repository.
Role	Chief Data Scientist/Business Analyst
Tasks executed	● Neural net model written in python using PyTorch framework, predicting the class of repo with avg accuracy of 75%.
Technologies used	Python, Mongo, GitHub

Project #8

Name of the project	**Marketing mix modeling for an advertising agency**
Description	Identify an optimal distribution of investments between various advertising channels like TV, Radio, Internet. The final model had to work efficiently for different products and brands of various clients of the agency.
Tasks executed	● Singular spectrum analysis for time series smoothing and future trend prediction ● Partial least squares (PLS) regression to get the value of variables impact ● Non-linear optimization to get optimal resources allocation ● Established reliable functions to explain nonlinear dependencies between variables ● Developed a reliable and accurate PLS regression model for preprocessed data
Technologies used	Python (sklearn, scipy, numpy, pandas)

Project #9

Name of the project	**Churn prediction model for a telecom company**
Description	Implement a mathematical model with automatically gathered data from various sources with the ability to give as an output a probability of churn for each client in the upcoming month.

Tasks executed	● Big data analysis of different sources: billing, network data, crm, client location info, mnp, call center data ● Implementation of a new architecture for data collection and Preprocessing ● Developing of ETL process for gathering necessary data for the model ● Classification algorithms based on python sklearn: Log Regression, SVM, Decision Trees, Random Forest, Gradient Boosting Algorithms ● Clustering based on python sklearn built-in algorithms
Technologies used	Python, Oracle RDBMS (main data storage), Cloudera, MySQL

Project #10

Name of the project	**Scoring model for a telecom company**
Description	The goal of the project was to predict the client's default and not to sell obligatory devices to unreliable clients (future debtors). The output of the model was a probability of being a debtor in a shortcoming perspective. Raw data we collected from Oracle RDBMS. Data analysis was performed using python pandas, numpy, re, matplotlib, seaborn, further data preprocessing and ETL was made on pl/sql. The mathematical model was developed using python imblearn & sklearn. The best results in terms of database integration velocity and accuracy were obtained using Random Forest. The results were also used for segmentation based on the client's financial status and paying discipline.
Role	Chief Data Scientist/Business Analyst/Project Manager
Tasks executed	● Big data analysis of different sources: billing, network data, crm, mnp, clients' photos, National Bank credit info data ● Implementation of a new architecture of data collection and preprocessing ● Creation of GUI interface for the scoring machine in the CRM system for sellers ● Validation test of several algorithms: log regression, decision trees, SVM, random forest, gradient boosting (chosen the best one) ● Developing ETL process for gathering necessary data for the model
Technologies used	Python, Oracle RDBMS (main data storage), MS Office, Windows

Project #11

Name of the project	**Multiclass classification model for an e-commerce company**

Description	The goal of the project was to predict the probability of purchase for each client of an e-commerce company. Raw data was collected using Logs API Yandex Metrics and analyzed in python using numpy, pandas. For storage and aggregation purposes we developed Clickhouse DB in cloud on Microsoft Azure under Ubuntu 16.04. Using python clickhouse module we preprocessed all the data and developed ETL process for data processing on a daily basis. The mathematical model was developed using python imblearn & sklearn. Algorithms we used for classification on a validation stage: Log Regression, SVM, Decision Trees, Random Forest, Gradient Boosting Algorithms. We divided a set of clients with predicted probability into 3 groups (excellent, good, normal) and sent them as segments to the advertisement platform using API for to retarget them and to attract new users using look-alike ML technology to increase conversion rate and decrease marketing campaign costs.
Role	Data Scientist/Business Analyst
Tasks executed	● Automatic data exchange with Logs API ● Implementation of data collection and preprocessing using Clickhouse in the cloud under Ubuntu ● Math model developing using python with algorithms validation checking procedure
Technologies used	Python, Clickhouse DB (main data storage), Ubuntu

Project #12

Name of the project	**DWH transformation and modernization for a telecom company**
Description	The goal of the project was an improvement of the ETL process, increasing the quality, accuracy, and velocity of the data collection, creating a new data model and data marts for BI tool and adding data from new sources with the enhancement of current attributes and metrics for the reporting purposes.
Role	Data Analyst/Business Analyst
Tasks executed	● Implementation of a new architecture for data collection ● Creation of new tables and views in Oracle RDBMS and new data marts for Microstrategy to meet the requirements of the business ● Increasing the velocity of data collection during ETL process ● Creation of new reports and dashboards in Microstrategy based on DWH data ● Checking system for possible DWH bugs and defects ● Update of Microstrategy BI to a newer version
Technologies used	Python, Microstrategy, Oracle RDBMS

Project #13

Name of the project	**Descriptive analysis for a telecom company**

Description	The project goal was to identify a typical client of the company and the patterns of the clients' behavior. The analysis was divided into several main parts: overall analysis with different dimensions, zero ARPU clients' analysis, data traffic usage analysis.
Role	Data Scientist/Business Analyst/Architect
Tasks executed	● The overall analysis of the data in different dimensions: which data we have, what is the level of their accuracy, where are they stored ● Creation of a dictionary of features that explained clients behavior with links to corresponding DBs ● Zero ARPU analysis using python pandas, numpy, matplotlib, seaborn with recommendations regarding how to decrease the number of such clients ● Data usage analysis using python pandas, numpy, matplotlib, seaborn ● Analysis of necessary database architecture improvement
Technologies used	Python, Oracle RDBMS, MySQL, MS Office, Windows

Project #14

Name of the project	**Call center performance analytical platform for a telecom company**
Description	The project goal was to develop a set of KPI for the call center and then create a system of online updated dashboards using the data in several database sources (Oracle, SQL Server, MySQL). Dashboards were implemented using Microstrategy 10 with creation of attributes and metrics and overall warehouse catalog improvement.
Role	Business Analyst/Microstrategy Architect
Tasks executed	● Analysis of several data sources (Oracle, SQL Server, MySQL) ● Creation of connections from Microstrategy to necessary DBs ● Creation of attributes, metrics and OLAP cubes in Microstrategy for implementation of dashboards ● Implementation of dashboards using Microstrategy 10 in the customer's native language
Technologies used	Microstrategy, Oracle RDBMS, SQL Server, MySQL, Windows

Project #15

Name of the project	**Financial assets forecasting**
Description	The goal of the project was to predict financial assets of the bank for accurate forecasting of the business plan KPI. Data analysis was developed in python pandas, numpy, matplotlib and seaborn, math model was developed using matlab. Due to high quasi-periodicity of data and their chaotic behavior, it was decided to use for prediction an algorithm singular spectrum analysis with principal component analysis to decrease the dimensionality of data series and smooth the data.

Role	Data Scientist/Business Analyst
Tasks executed	● Developing a mathematical model for assets prediction ● Implementation of a new business plan template with an integrated model in it
Technologies used	Python, Matlab