# Satwant Singh

**In**: https://in.linkedin.com/in/satwant-singh201
**Portfolio:** https://satwant-ds.github.io/Satwant-DS

satwants@usc.edu
(213) 326-8089
Los Angeles, CA, 90007

## EDUCATION

| | |
|---|---|
| **University of Southern California** | Los Angeles |
| Master of Science in Applied Data Science | **AUG 2021-MAY 2023** |

**Relevant Coursework:** Foundations of Data Management, Machine Learning for Data Science

| | |
|---|---|
| **Punjab Engineering College** | Chandigarh |
| Bachelor of Engineering in Electronics & Electrical Communication - **GPA: 9.4/10** | **AUG 2013-MAY 2017** |

**Achievements:** TCS Gold Medal for Best Capstone project 2017 (**Brain Computer Interfacing**)

## SKILLS

- **Languages**: Python, Apache Hive, PySpark, SQL, Basic HTML & CSS, Shell Scripting
- **Data Science Skills**: Linear Regression, Random Forest, GBM, Clustering, CHAID Analysis, Word Embeddings, Map Reduce, Spark, Sequence Modeling, H2o, TFIDF, Recommender Systems, Risk Analytics, Hadoop, Git, Unix, MS Excel, MS Excel Solver
- **Data Science Libraries**: Pandas, H2o, Lime, NumPy, SciPy, Scikit-Learn, NLTK, Gensim, Fastai, Keras, Spark, Matplotlib, seaborn
- **Databases**: MS SQL Server, DynamoDB, Firebase, AWS S3, MongoDB, MySQL, HBase, HDFS

## WORK EXPERIENCE

| | |
|---|---|
| **UnitedHealth Group – Center for Advanced Research and Analytics** | Gurugram, India |
| **Data Scientist (Project Lead)** | **AUG 2019-JUL 2021** |
| **Non-Clinical Audit Prioritization Algorithm** | **NOV 2019-JUL 2021** |

- Designed Predictive Model with 95% Precision aimed at reducing False Positive Rate for Fraud detection Pre-Pay Audit Process
- Deployed PySpark Application with 2 GBM models & 1 GLM model object on TWS with HBase based archival mechanism
- Reduced FPR for client team by 4% and increased Audit driven Savings by 3M USD per Month
- Standardized Audit process enabled Client to lower audit dollar threshold by 20% and helped identify new error categories
- Delivered 4 Tableau dashboards to Client & ARA Leadership for project performance tracking & conducting cost benefit analysis
- Expanded project to all 3 Lines of Business in UHG and overlooked model development for 2nd generation models

**Medical Fraud Risk Authentication**                                                                                    **APR 2019-SEP 2020**

- Established a new opportunity of 10% reduction in clinical investigation volumes by proposing a semi supervised ML solution
- Mined new error patterns by building clinical risk prediction model; Automated 2 error categories with 100% True positive rate using SQL Jobs; Created binary classification ML model for 4 value streams within 100M USD Payment Integrity Portfolio
- Collaborated with Client to identify new fraud patterns and helped reduce FPR for overall Clinical Investigation by **7%** and enabled reduction in administrative costing of **4M USD per month**
- Served HHS.gov in helping them capture Covid-19 RT PCR testing fraud and reported top 10 contributing medical facilities

| | |
|---|---|
| **UnitedHealth Group – Process Optimization Analytics Team** | Noida, India |
| **Associate Data Scientist** | **JUN 2017-JUL 2019** |
| **AutoML Toolkit** | **JAN 2019-MAR 2019** |

- Orchestrated PySpark Package for creating 4 Feature types in Risk Analytics namely Look-back, Risk, Entropy and Risk Severity
- Scripted stable version of Toolkit using PySpark & SQL utilized by 13 Data Science Teams across 20+ machine learning projects
- Achieved 90% reduction in Model development time frame by automating Data Preparation & Feature Engineering modules

**Overpayment Detection Engine**                                                                                         **APR 2018-DEC 2018**

- Migrated Logistic Regression Model built in SAS to PySpark achieving run time reduction from 7 days to 30 mins
- Gathered knowledge on insurance claim life cycle and identified key driving business factors impacting claim cost
- Identified true pricing labels for 100M records big sample using 8 data bases and prepared analytical dataset for modelling
- Built H2o GLM Model with adjusted R2 of 89% with Lasso regularization and several data transformations
- Setup weekly Lead release mechanism in Python for sharing top 100 Doctors and published results on Oracle SQL Server
- Conducted regular Look-back feature refresh every 30 days for capturing seasonal trends in pricing

**Next Best Provider Identification**                                                                                      **SEP 2017-MAR 2018**

- Evaluated network adequacy recommendations by CMS.gov and identified counties with <90% system adequacy in terms of patient-doctor geographical distance for all members across specialties & states
- Leveraged Knapsack algorithm and Vincenty's formulae for identifying new doctors capable of catering unserved members
- Executed 100% ETL in Apache Hive and Shell Scripting and generated Provider recommendations for UHN marketing teams