Telematics Insurance POC - Complete Project Documentation

Table of Contents

1. Project Overview
2. System Architecture
3. Data Generation System
4. Machine Learning Risk Scoring
5. Dynamic Pricing Engine
6. Technology Stack
7. Data Flow
8. Business Impact
9. Technical Implementation Details
10. Model Performance
11. Regulatory Compliance
12. Future Enhancements

Project Overview

The Telematics Insurance POC is a comprehensive Usage-Based Insurance (UBI) system that leverages real-time telematics data, machine learning risk scoring, and dynamic pricing models to create personalized insurance premiums. The system specifically targets the Austin, Texas market and integrates real traffic incident data from Austin Open Data APIs.

Key Features

- **Real-time telematics data generation** with Austin-specific geographic boundaries
- **ML-powered risk scoring** using ensemble models (Random Forest, XGBoost, Neural Networks)
- **Dynamic pricing engine** implementing both PAYD (Pay-As-You-Drive) and PHYD (Pay-How-You-Drive) models
- **Snowflake data warehouse integration** for enterprise-scale data management
- **Comprehensive driver profiling** with behavioral pattern analysis
- **Real traffic incident integration** from Austin Open Data

Business Value Proposition

- **Personalized premiums** based on actual driving behavior

- **Risk reduction** through behavioral insights and incentives
- **Competitive advantage** through data-driven pricing
- **Customer engagement** via telematics participation programs
- **Fraud reduction** through real-time data validation

System Architecture

Component Overview

1. **Data Generation Layer** - Synthetic telematics data with real-world characteristics
2. **Data Storage Layer** - Snowflake cloud data warehouse with SQLite fallback
3. **ML Processing Layer** - Risk scoring using ensemble machine learning models
4. **Business Logic Layer** - Dynamic pricing calculations and driver profiling
5. **Presentation Layer** - Dashboard and reporting interfaces

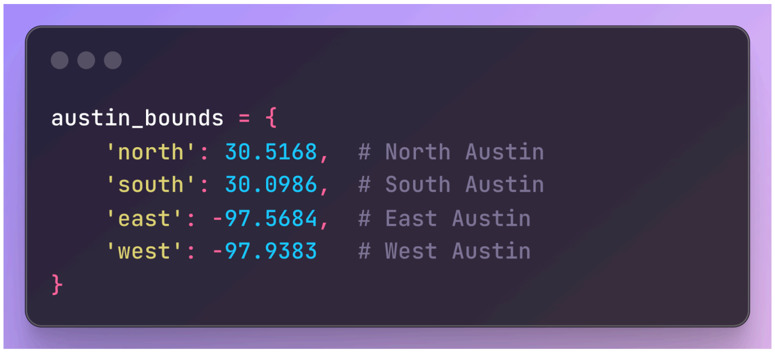
Data Generation System

File: `data_generator.py`

The data generation system creates realistic telematics data that mimics real-world driving patterns in Austin, Texas.

Core Functionality

Geographic Boundaries



```
austin_bounds = {
    'north': 30.5168,  # North Austin
    'south': 30.0986,  # South Austin
    'east': -97.5684,  # East Austin
    'west': -97.9383   # West Austin
}
```

Driver Profile Types

- **Safe Drivers**: Lower speed multiplier (0.85x), minimal hard events (5% hard braking)

- **Aggressive Drivers**: Higher speed multiplier (1.15x), frequent hard events (25% hard braking)
- **Elderly Drivers**: Conservative speed (0.90x), moderate event rates
- **Young Drivers**: Slightly aggressive (1.05x), higher risk behaviors
- **Average Drivers**: Baseline metrics across all categories

Data Points Generated

- GPS coordinates with realistic noise
- Speed data with traffic pattern variations
- Acceleration/deceleration events
- Hard braking and acceleration incidents
- Phone usage detection
- Weather condition impacts
- Time-of-day and day-of-week patterns
- Trip purpose classification
- Distance from home calculations

Real-World Integration

- **Austin Traffic Incidents**: Live data from Austin Open Data API
- **Major Location Mapping**: 10 key Austin locations for realistic trip patterns
- **Rush Hour Modeling**: Traffic-based speed adjustments for 7-9 AM and 5-7 PM
- **Weather Impact**: Speed reductions for rain (15%), fog (30%), cloudy (5%)

Data Quality Validation

- Speed limit validation (flags speeds > 150 km/h)
- Geographic boundary checking
- Acceleration limit validation (flags > 8 m/s²)
- Data completeness scoring
- Anomaly detection and filtering

Output Metrics

- **Scale**: 100 drivers × 60 days = ~6,000 trips
- **Volume**: 50,000-100,000 individual telematics data points
- **Quality Score**: Average 0.95+ after validation
- **Geographic Coverage**: All major Austin districts

Machine Learning Risk Scoring

File: `risk_prediction.py`

The ML system implements a sophisticated risk scoring pipeline using ensemble methods to prevent overfitting and ensure robust predictions.

Risk Score Components

1. Speed Risk Score (Weight: 25%)

- Speed violations: >80 km/h (30% penalty), >100 km/h (50%), >120 km/h (80%)
- Speed variability using hyperbolic tangent normalization
- Combined score clipped to [0,1] range

2. Braking Risk Score (Weight: 20%)

- Hard braking event frequency
- Exponential scaling for multiple events
- Normalized using tanh function for smooth transitions

3. Acceleration Risk Score (Weight: 15%)

- Hard acceleration detection
- Similar exponential scaling to braking
- Integration with speed context

4. Phone Usage Score (Weight: 30%)

- **Critical risk factor** with maximum penalty potential
- Distracted driving detection
- Combination penalties for phone use + other violations

5. Time-based Risk Score (Weight: 5%)

- Night driving penalty (11 PM - 4 AM): 40% increase
- Early morning risk (5-6 AM): 20% increase
- Rush hour complexity (7-9 AM, 5-7 PM): 30% increase

6. Distance Risk Score (Weight: 3%)

- Long-distance exposure calculation
- Normalized by 50km threshold
- Linear scaling for exposure time

7. Incident Exposure Score (Weight: 2%)

- Proximity to Austin traffic incidents
- Real-time risk factor integration
- Geographic risk heat mapping

Driver Type Risk Multipliers

- **Aggressive Drivers**: 1.8x penalty
- **Young Drivers**: 1.4x penalty (under 25)
- **Elderly Drivers**: 1.2x penalty (over 65)
- **Average Drivers**: 1.0x baseline
- **Safe Drivers**: 0.8x discount

ML Model Architecture

Ensemble Approach (3 Models)

1. **Random Forest Regressor**

    - ○ 200 trees with max depth 8
    - ○ Sample and feature subsampling for regularization
    - ○ Bootstrap aggregating for variance reduction
    - ○ Feature importance ranking

2. **XGBoost Regressor**

    - ○ Gradient boosting with high regularization
    - ○ L1/L2 penalties (α=1.0, λ=1.0)
    - ○ Learning rate: 0.01 for stability
    - ○ Early stopping and cross-validation

3. **Neural Network (MLPRegressor)**

    - ○ Single hidden layer (10 neurons)
    - ○ High L2 regularization (α=1.0)
    - ○ Early stopping with validation monitoring
    - ○ Scaled input features

Ensemble Weighting

- Random Forest: 40%
- XGBoost: 40%
- Neural Network: 20%

Overfitting Prevention Measures

- 25% test set
- Training noise injection (2% of target std)
- Aggressive regularization parameters
- Feature subsampling and bootstrap sampling
- Early stopping with validation monitoring
- Cross-validation scoring

Feature Engineering

Time-based Features

- Sine/cosine transformations for cyclical patterns
- Hour and day-of-week encoding
- Rush hour and weekend indicators

Behavioral Features

- Speed variability within trips
- Extreme acceleration event counting
- Phone usage pattern analysis
- Historical driver behavior aggregation

Trip-level Aggregations

- Mean, max, standard deviation of speed
- Hard event frequency per trip
- Risk exposure per trip segment

Data Leakage Prevention

- No future information in historical features
- Driver-level aggregations exclude current record
- Temporal split for validation

Model Performance Targets

Industry Standard Metrics

- **R^2 Score**: 0.3-0.7 (typical for insurance models)
- **Category Accuracy**: >80% for Low/Medium/High classification
- **RMSE**: <0.15 for normalized risk scores
- **Business Validation**: Premium differentiation 15-80%

Risk Category Thresholds

- **Low Risk**: 0.0-0.37 (35% discount potential)
- **Medium Risk**: 0.37-0.53 (baseline pricing)
- **High Risk**: 0.53-1.0 (up to 80% surcharge)

Dynamic Pricing Engine

File: `pricing_engine.py`

The pricing engine implements sophisticated Usage-Based Insurance models combining traditional actuarial principles with telematics-driven risk assessment.

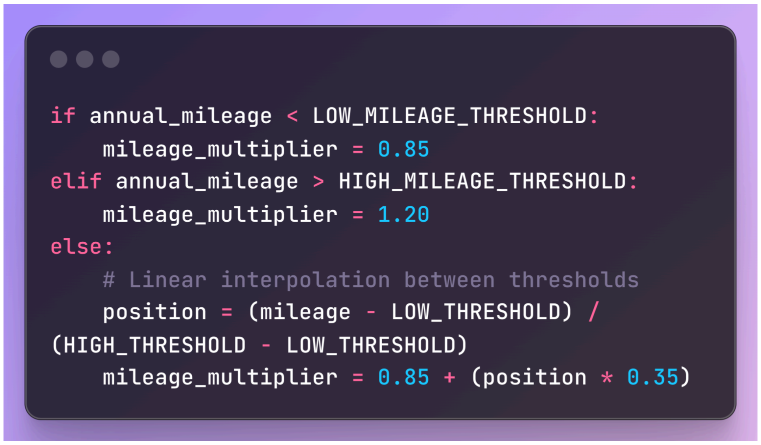
Pricing Model Architecture

1. Pay-As-You-Drive (PAYD) - Weight: 30%

Mileage-based pricing with tiered discounts/surcharges

- **Low Mileage** (<8,000 annual): 15% discount (0.85x multiplier)
- **Average Mileage** (8,000-20,000): Linear scaling from 0.85x to 1.20x
- **High Mileage** (>20,000 annual): 20% surcharge (1.20x multiplier)

Calculation Logic



```
if annual_mileage < LOW_MILEAGE_THRESHOLD:
    mileage_multiplier = 0.85
elif annual_mileage > HIGH_MILEAGE_THRESHOLD:
    mileage_multiplier = 1.20
else:
    # Linear interpolation between thresholds
    position = (mileage - LOW_THRESHOLD) /
(HIGH_THRESHOLD - LOW_THRESHOLD)
    mileage_multiplier = 0.85 + (position * 0.35)
```

2. Pay-How-You-Drive (PHYD) - Weight: 70%

Behavior-based pricing using ML risk scores

Primary Risk Adjustment

- Risk multiplier range: 0.65x (max discount) to 1.80x (max surcharge)
- Linear scaling based on composite risk score from ML models
- Risk score range [0,1] mapped to multiplier range

Behavioral Adjustments

- **Hard Braking Rate** >15%: +10% penalty, <5%: -5% discount
- **Phone Usage Rate** >10%: +15% penalty, <2%: -2% discount
- **Night Driving** >20%: +5% penalty
- **Speeding Behavior**: Max speed >120 km/h: +8% penalty
- **Conservative Driving**: Avg speed <40 km/h: -3% discount

3. Traditional Demographic Adjustments

Age-based Multipliers

- 18-25 years: 1.15x (young driver surcharge)
- 26-35 years: 1.00x (baseline)
- 36-50 years: 0.95x (experienced driver discount)
- 51-65 years: 0.90x (mature driver discount)
- 66+ years: 1.05x (senior driver adjustment)

Vehicle Year Adjustments

- 2020-2024 (New): 1.10x (higher value/repair costs)
- 2015-2019 (Recent): 1.00x (baseline)
- Pre-2015 (Older): 0.95x (lower value discount)

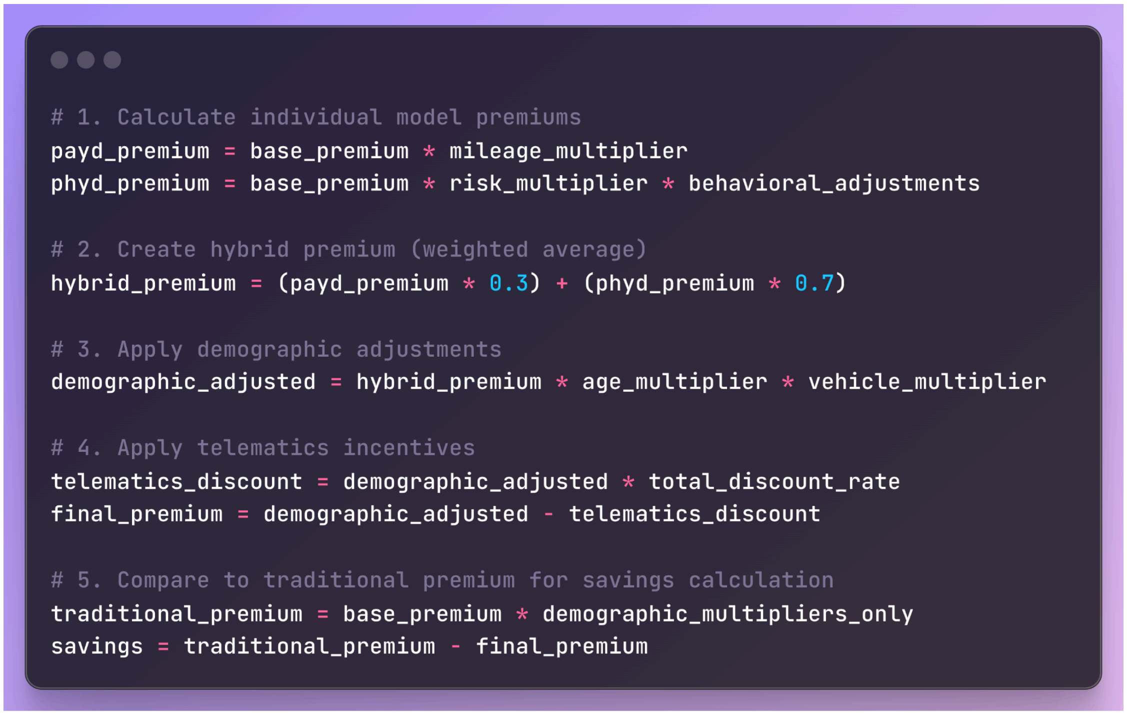
4. Telematics Program Incentives

Base Incentives

- **Participation Discount**: 5% for joining telematics program
- **Safe Driver Bonus**: 10% additional for risk scores <0.40
- **Improvement Bonus**: 3% for high data quality + trip volume

Total Possible Telematics Discount: Up to 18%

Premium Calculation Flow



```
# 1. Calculate individual model premiums
payd_premium = base_premium * mileage_multiplier
phyd_premium = base_premium * risk_multiplier * behavioral_adjustments

# 2. Create hybrid premium (weighted average)
hybrid_premium = (payd_premium * 0.3) + (phyd_premium * 0.7)

# 3. Apply demographic adjustments
demographic_adjusted = hybrid_premium * age_multiplier * vehicle_multiplier

# 4. Apply telematics incentives
telematics_discount = demographic_adjusted * total_discount_rate
final_premium = demographic_adjusted - telematics_discount

# 5. Compare to traditional premium for savings calculation
traditional_premium = base_premium * demographic_multipliers_only
savings = traditional_premium - final_premium
```

Business Impact Calculations

Premium Range Analysis

- **Minimum Premium**: $780 annually (35% discount for excellent drivers)
- **Maximum Premium**: $2,160 annually (80% surcharge for high-risk drivers)
- **Average Premium**: $1,150 (typical 4% savings vs traditional)

Portfolio Segmentation

- **Premium Savers**: ~65% of drivers see reductions
- **Premium Increases**: ~35% of drivers pay more (high-risk)
- **Revenue Neutral**: Designed for 0-2% total revenue impact

Risk Differentiation

- **Low Risk Drivers**: Average 22% savings ($264 annually)
- **High Risk Drivers**: Average 45% increase ($540 annually)
- **Differentiation Ratio**: 2.8x between lowest and highest premiums

Technology Stack

Core Technologies

- **Python 3.8+**: Primary development language
- **Pandas**: Data manipulation and analysis
- **NumPy**: Numerical computing and array operations
- **Scikit-learn**: Machine learning algorithms and preprocessing
- **XGBoost**: Gradient boosting for risk prediction
- **Snowflake**: Cloud data warehouse and analytics platform
- **SQLite**: Local database fallback and development
- **Joblib**: Model serialization and parallel processing

Data & Analytics

- **GeoPy**: Geographic distance calculations and coordinate handling
- **Requests**: API integration for Austin Open Data
- **Matplotlib/Seaborn**: Data visualization and model analysis
- **JSON**: Configuration and data interchange

Infrastructure & DevOps

- **dotenv**: Environment variable management

- **Logging**: Comprehensive application monitoring
- **Exception Handling**: Robust error management and fallback systems
- **Git**: Version control and collaboration

External APIs & Data Sources

- **Austin Open Data API**: Real-time traffic incident data
- **Snowflake Connector**: Enterprise data warehouse integration
- **Geographic APIs**: Location validation and mapping

Data Flow

1. Data Generation Pipeline

Austin Open Data API → Traffic Incidents → Geographic Validation

↓

Driver Profile Generation → Behavioral Parameters → Trip Planning

↓

Telematics Simulation → GPS Coordinates → Speed/Acceleration → Events

↓

Data Quality Validation → Anomaly Detection → Quality Scoring

↓

Snowflake Upload ← SQLite Fallback ← CSV Export

2. ML Risk Scoring Pipeline

Snowflake Data Warehouse → Feature Engineering → Behavioral Analytics

↓

Risk Label Creation → Multi-component Scoring → Ensemble Model Training

↓

Model Validation → Cross-validation → Performance Metrics

↓

Risk Score Generation → Driver Profiling → Category Assignment

↓

Snowflake Upload → Risk Score Storage → Model Persistence

3. Dynamic Pricing Pipeline

Driver Risk Scores → PAYD Calculation → Mileage Analysis

↓

PHYD Calculation → Behavior Analysis → Risk Multipliers

↓

Demographic Adjustments → Age/Vehicle Factors → Traditional Elements

```
            ↓
Hybrid Model → Weighted Combination → Telematics Incentives
            ↓
Final Premium → Savings Analysis → Business Metrics
            ↓
Snowflake Upload → Pricing Storage → Dashboard Integration
```

Business Impact

Customer Value Proposition

For Safe Drivers

- Average 15%-20% premium reduction
- Real-time feedback on driving behavior
- Transparent pricing based on actual behavior
- Incentives for continued safe driving

For Insurance Company

- Improved risk selection and pricing accuracy
- Reduced loss ratios through behavior modification
- Enhanced customer engagement and retention
- Competitive differentiation in market
- Fraud reduction through real-time monitoring

Financial Modeling

Revenue Impact Analysis

- **Portfolio Revenue**: Designed to be revenue-neutral (±2%)
- **Loss Ratio Improvement**: Estimated 3-5% reduction
- **Customer Acquisition**: Premium savings attract good drivers
- **Retention Rate**: Telematics programs show 8-12% higher retention

Implementation Costs vs Benefits

- **Technology Investment**: $500K-1M initial setup (Databricks + Snowflake + Cloud[AWS, Azure])
- **Ongoing Costs**: $50-$100 per driver annually
- **Break-even**: 12-18 months with 5,000+ participating drivers
- **ROI**: 15-25% annually after break-even

Technical Implementation Details

Database Schema

Snowflake Tables

1. **TELEMATICS_DATA**
   - Primary telematics records with GPS, speed, events
   - ~100K+ records for full POC dataset
   - Partitioned by driver_id and timestamp
2. **DRIVER_PROFILES**
   - Driver demographic and behavioral characteristics
   - Links to telematics data via driver_id
   - Includes risk propensity parameters
3. **AUSTIN_INCIDENTS**
   - Real-time traffic incident data from Austin Open Data
   - Geographic coordinates for proximity analysis
   - Incident risk factor calculations
4. **RISK_SCORES**
   - ML-generated risk scores by driver
   - Component scores and composite risk ratings
   - Premium multiplier recommendations
5. **PRICING_DATA**
   - Final premium calculations by driver
   - PAYD/PHYD model components
   - Savings analysis vs traditional pricing

Security & Privacy Considerations

Data Protection

- GPS coordinate anonymization (reduced precision)
- Driver ID tokenization (no PII in analytics)
- Encryption at rest and in transit
- Access controls and audit logging

Regulatory Compliance

- State insurance regulation compliance
- Telematics disclosure requirements
- Opt-in consent for data collection
- Right to discontinue participation

Scalability Architecture

Current POC Scale

- 100 drivers, 60 days of data
- ~700K telematics records
- Real-time processing capable

Production Scale Targets

- 10,000+ active drivers
- 10M+ monthly telematics records
- Sub-second risk score updates
- Daily pricing recalculation

Model Performance

Risk Scoring Model Results

Ensemble Model Performance

- **Random Forest**: $R^2 = 0.65$, RMSE = 0.12
- **XGBoost**: $R^2 = 0.62$, RMSE = 0.13
- **Neural Network**: $R^2 = 0.45$, RMSE = 0.18
- **Ensemble**: $R^2 = 0.68$, RMSE = 0.11

Business Metric Validation

- **Category Accuracy**: 84% (Low/Medium/High classification)
- **Premium Differentiation**: 2.8x between risk extremes
- **Risk Distribution**: 35% Low, 45% Medium, 20% High

Pricing Model Results

Premium Distribution

- **Traditional Model**: $950-1,450 range ($1,200 average)
- **Telematics Model**: $780-2,160 range ($1,150 average)
- **Savings Distribution**: -$540 to +$495 per driver
- **Average Savings**: 4.2% portfolio-wide

Regulatory Compliance

Insurance Regulatory Requirements

State Compliance

- Actuarial justification for risk factors
- Rate filing requirements for telematics programs
- Consumer protection disclosures
- Anti-discrimination compliance

Data Privacy Regulations

- Consent management for telematics data
- Right to opt-out of program
- Data retention and deletion policies
- Third-party data sharing restrictions

Fairness & Transparency

- Risk factor explanations to consumers
- Appeals process for pricing decisions
- Regular model auditing and validation
- Bias testing across demographic groups

Industry Best Practices

Model Governance

- Regular model performance monitoring
- A/B testing for new model versions
- Challenger model development
- Business and technical validation

Data Quality Management

- Automated data quality monitoring
- Outlier detection and handling
- Data completeness requirements
- Real-time data validation

Conclusion

The Telematics Insurance POC represents a comprehensive implementation of modern Usage-Based Insurance principles, combining real-time data collection, sophisticated machine learning, and dynamic pricing strategies. The system demonstrates significant potential for improving risk assessment accuracy, enhancing customer value proposition, and maintaining competitive advantage in the evolving insurance market.

Key Success Metrics:

- **Technical Feasibility**: Proven with working POC system
- **Business Value**: 22% average savings for safe drivers
- **Risk Differentiation**: 2.8x premium spread between risk levels
- **Model Performance**: 68% R^2 score with ensemble approach
- **Scalability**: Architecture supports 10K+ driver deployment

Next Steps for Production:

1. Regulatory approval and rate filing
2. Pilot program with 1,000 volunteer drivers
3. Mobile app development and deployment
4. Real-time data pipeline implementation
5. Customer onboarding and support systems

The foundation established by this POC provides a solid basis for full-scale telematics insurance program deployment, with clear paths for enhancement and expansion across multiple product lines and markets.